Bayesian nonparametric spatial prior for traffic crash risk mapping: a case study of Victoria, Australia

Jean-Baptiste Durand, Florence Forbes, Cong Duc Phan, Long Truong, Hien D Nguyen, Fatoumata Dama

Aust. N. Z. J. Stat. 2020 doi: 10.1111/j.1467-842X.XXX

Bayesian nonparametric spatial prior for traffic crash risk mapping: a case study of Victoria, Australia

J.-B. Durand[1*], F. Forbes[1], C.D. Phan[2], L. Truong[2], H.D. Nguyen[2,3], and F. Dama[1]

Inria, LJK, Statify Team, La Trobe University, University of Queensland

Summary

We investigate the use of Bayesian nonparametric (BNP) models coupled with Markov random fields (MRF) in a risk mapping context, to build partitions of the risk into homogeneous spatial regions. In contrast to most existing methods, the proposed approach does not require an arbitrary commitment to a specified number of risk classes and determines their risk levels automatically. We consider settings in which the relevant information are counts and propose a so-called BNP Hidden MRF (BNP-HMRF) model that is able to handle such data. The model inference is carried out using a variational Bayes Expectation–Maximisation algorithm and the approach is illustrated on traffic crash data in the state of Victoria, Australia. The obtained results corroborate well with the traffic safety literature. More generally, the model presented here for risk mapping offers an effective, convenient and fast way to conduct partition of spatially localised count data.

Key words: Road safety; Traffic crashes; Risk mapping; Bayesian nonparametrics; Markov random field; Variational Bayes Expectation–Maximisation algorithm

1. Introduction

Traffic-related injuries and deaths are major problems in contemporary societies. Social economic losses from traffic crashes, in particular from motor vehicle crashes, can be enormous. This makes road and traffic safety a major concern, worldwide. The nondecreasing relationship between crash casualties and population suggests that safety improvements could be gained from a better prediction of crash occurrences. Traffic crashes are complex events involving the interactions of various factors. In particular, since road transport involves distances by nature, most studies call for spatial analysis to account for geographical locations

*Author to whom correspondence should be addressed.
[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Inria Grenoble Rhone-Alpes, 655 av. de l'Europe, 38335 Montbonnot, France
[2] School of Engineering and Mathematical Sciences, La Trobe University, Bundoora, Australia
[3] School of Mathematics and Physics, University of Queensland, St. Lucia, Australia
Email: Jean-Baptiste.Durand@inria.fr

Acknowledgment. The authors are partly supported by the Inria project Lander.

and environments in which crashes occur. The goal is often to accurately predict the risks at different locations (Lord & Mannering 2010) and to link these risk values to other variables for interpretability, or to assess the impact of several risk factors (Theofilatos & Yannis 2014; Papadimitriou et al. 2019) and road safety measures (Elvik et al. 2009). The hope is to identify the potential causal sources of crashes, and to apply appropriate control procedures and protection measures; see e.g., Truong & Currie (2019).

In this work, our goal is not to predict the number of crashes or the related risk, as per se, although predictions are an output of our model. We primarily aim at highlighting areas with different risk levels, with respect to various covariates, such as population density, traffic density, signalisation density, etc. The interest of such partitioning is to highlight spatial heterogeneity, to locate high risk areas (so-called risk hot spots), and to determine whether the data exhibit some structure in space that could be analysed or directly interpreted. Moreover, aggregation of connected regions with low numbers of crashes and the same risk level can also be used to increase the effective sample size dedicated to estimating that risk level.

Such partitions can be obtained by applying risk mapping models. Standard risk mapping models usually produce a continuous estimation of the risk that requires a post-processing classification step to obtain clearly delimited risk zones, usually based on an arbitrary choice of risk levels or of risk intervals. Examples of ways to set these thresholds are illustrated in the Crash Risk Mapping Technical Specifications report of the European Road Assessment Programme: `https://eurorap.org/crash-rate-mapping/`.

Most statistical methods for risk mapping of aggregated data (e.g., Mollié 1999; Richardson et al. 1995; Pascutto et al. 2000; Lawson et al. 2000) are based on a Poisson log-linear mixed model and follow the so-called BYM model of Besag, York & Mollié (1991), and extended by Clayton & Bernadinelli (1992), which is called the convolution model by Mollié (1996). This model is based on a Hidden Markov Random Field (HMRF), where the latent intrinsic risk field is modeled by a Markov field with continuous state space, namely a Gaussian Conditionally Auto-Regressive (CAR) model. Other developments in this context concern spatio-temporal mapping (Knorr-Held & Richardson 2003; Robertson et al. 2010; Lawson & Song 2010) and multivariate risk mapping (Knorr-Held, Rasser & Becker 2002; MacNab 2010).

For all of these procedures, the model inference results in a real-valued estimation of the risk at each location and one of the main reported limitations is that local discontinuities in the risk field are not modelled (see e.g., Green & Richardson 2002), leading to potentially oversmoothed risk maps. Also, in some cases, coarser representations where areas with similar risk values are grouped are desirable (e.g., Abrial et al. 2005). Grouped representations have the advantage of providing clearly delimited areas in which more focused studies could be conducted to better understand the crashes determinants. These areas

at risk can be viewed as clusters as in Knorr-Held & Rasser (2000), but we prefer to interpret them as risk classes, as per the seminal work of Schlattmann & Böhning (1993) and Böhning, Dietz & Schlattmann (2000), and with additional spatial constraints, by Green & Richardson (2002) and Alfo, Nieddu & Vicari (2009).

Indeed, geographically separated areas representing different clusters can have similar risks and be grouped in the same class. Consequently, the classes can be less numerous than the number of clusters, and their interpretation is made easier for decision-makers. Using the BYM model, it is possible to derive such a grouping from the output, using either fixed risk ranges (usually difficult to choose in practice), or more automated clustering techniques (see e.g., Fraley & Raftery 2007). In any case, this post-processing step is likely to be sub-optimal. For this reason, there have been several attempts to design procedures that can directly model such a risk classification.

Green & Richardson (2002) propose replacing the continuous risk field by a partition model, involving the introduction of a finite number of risk levels and allocation variables to assign each area under study to one of these levels. Spatial dependencies are then taken into account by modeling the allocation variables as a latent discrete state-space Markov field, such as the Ising (two classes or states) or Potts (more than two classes) model (Chandler 1987; Stoehr 2017). In the same spirit, in the work by Fernandez & Green (2002), the spatial dependence is pushed one level higher, resulting in a more flexible model, but with a more difficult parameter estimation problem. These various attempts are based on discrete HMRF modelling and all use MCMC techniques for inference, which can seriously limit and even prevent their application to large data sets in a reasonable time.

An additional problem that arises in the application of HMRF is that of choosing the number of states, or classes. When data are independent, the typical approach is to use a penalised likelihood or information theoretic method, such via the Akaike information criterion, Bayesian information criterion, integrated likelihood, or other similar approaches. We note that some of these approaches have been extended to the HMRF framework (see, e.g., Forbes & Peyrard 2003). A drawback of such approaches is that they are computationally wasteful. That is, one is required to estimate models corresponding to different numbers of classes, and only the best of these models is then used for further inference. In order to alleviate this shortcoming, fully Bayesian methods such as the reversible jump Markov chain Monte Carlo (Green 1995; Miller & Harrison 2018) may be considered, although such procedures may actually require a greater amount of overall computational effort, regardless of wastefulness.

In this work, to handle discontinuities in the spatial structure of the risk without having to arbitrarily choose their number of classes, we propose to operate in the framework of Bayesian nonparametric (BNP) methods (Ghosal & Van der Vaart 2017). More specifically,

we build on methods recently proposed for the modelling of continuous observations by Lü, Arbel & Forbes (2020). We extend the approach, referred to as BNP-HMRF, to the modelling of count data. We derive the corresponding variational Bayes Expectation–Maximisation (VBEM) algorithm for the model estimation. The approach is then illustrated on traffic crash data in the state of Victoria, Australia. The analysis provides risk zones and risk levels that are globally coherent with other findings in the literature.

The proposed BNP-HMRF model is explained in Section 2. The model implementation using variational approximation is detailed in Section 3. The application to crash risk mapping of Victoria is detailed in Section 4. An assessment of the Markov random field hyperparameter estimation based on simulated data is proposed in Section 5. A conclusion and perspectives are provided in Section 6.

2. BNP-HMRF model for count data

The study aims at providing some risk mapping of traffic crashes, based on data regarding geographical zones. Since, on average, the number of traffic crashes increases with respect to other variables characterising the traffic importance (e.g., population size, traffic intensity, length of road network), the numbers of crashes have to be normalised with respect to at least one of these variables. The obtained ratios provide quantities that we may interpret as risks. One objective is then to account for some spatial heterogeneity regarding the observed risks. The model described in the following lines aims at clustering regions with similar risks to provide a labelled map, where each label is associated with some risk level.

Consider J regions, where we let y_j represent the number of crashes occurring in region $j \in \{1, \dots, J\}$, characterised by a normalisation variable N_j, e.g., the population size of region j. For the sake of clarity, we first assume that there is a finite set of K risk levels $\mathbf{\Lambda} = \{\lambda_0, \dots, \lambda_{K-1}\}$, that are ordered so that λ_k is the $(k+1)\text{th}$ smallest level. Since the risk level associated with region j is not known in advance, a variable $z_j \in \{0, \dots, K-1\}$ is introduced to indicate the assigned risk level, i.e. $z_j = k$, when region j is at risk level λ_k. When region j is at risk level λ_k, the number of traffic crashes y_j is then assumed to be Poisson distributed with mean $\lambda_k N_j$. That is, y_j conditioned on $z_j = k$ has probability mass function

$$p(y_j | z_j = k; \mathbf{\Lambda}, N_j) = \mathcal{P}(y_j; \lambda_k N_j), \tag{1}$$

where $\mathcal{P}(\cdot; \lambda_k N_j)$ denotes the probability mass function of the Poisson distribution with mean parameter $\lambda_k N_j$. Generically, $p(y_j | z_j = k; \mathbf{\Lambda}, N_j)$ is referred to as an emission

distribution. As a consequence, the mean number of crashes is a linear function of N_j: $\mathrm{E}[y_j|z_j = k; \mathbf{\Lambda}, N_j] = \lambda_k N_j$.

The goal is then to estimate the risk levels $\mathbf{\Lambda} = \{\lambda_0, \ldots, \lambda_{K-1}\}$ and the most likely risk mapping of each region, through the most likely values of the variables z_js. In practice, risk levels are likely to vary smoothly across regions. It is more likely that neighbouring regions have the same risk level with possible abrupt changes from one region to another if they have contrasting characteristics. Thus, for a better estimation of risk levels, a Markov random field (MRF) model is used for the set of labels $\boldsymbol{z} = \{z_j, j \in J\}$, to account for spatial dependencies between connected regions.

Formally, the regions are seen as the vertices of a graph G. They are connected by an edge in the graph whenever they share a boundary, although other types of connections could be considered (e.g., they either share a boundary or have a common neighbour, etc.). The probability for neighbouring regions having either a similar or different label is controlled by some scalar positive parameter denoted by β. The higher the value of β, the more likely neighbouring regions are at the same risk level.

The number K of risk levels is not usually known in advance and has to be chosen adaptively by users. To avoid this commitment to a fixed number K, we propose an extension of the model that does not restrict the levels to a finite number K. This extension is based on so called Dirichlet Process Mixtures (Lü, Arbel & Forbes 2020) and is referred to as a Bayesian Non-Parametric Hidden Markov Random Field (or more concisely, BNP-HMRF).

The BNP-HMRF model is defined as follows. The set of J regions under consideration is associated to a graph structure $G = (J, E)$, where each $j \in J$ corresponds to a node of G and the set of edges E represents all pairs of regions with a common boundary. The likelihood part of the model is given by (1). The observations are counts $\boldsymbol{y} = \{y_j, j \in J\}$ distributed independently given $\mathbf{\Lambda}$ and $\boldsymbol{z}$ with for every j,

$$p(y_j|z_j = k; \mathbf{\Lambda}, N_j) = \mathcal{P}(y_j; \lambda_k N_j).$$

The risk class labels $\boldsymbol{z} = \{z_j, j \in J\}$ are assumed to be distributed as a Markov random field on G with the following distribution:

$$\begin{aligned} p(\boldsymbol{z}|\beta, \boldsymbol{\pi}) &\propto \exp\left(\sum_{j=1}^{J} \ln\, \pi_{z_j} + \beta \sum_{\{i,j\}\in E} \mathbf{1}_{(z_i=z_j)}\right) \\ &= \left(\prod_{j=1}^{J} \pi_{z_j}\right) \exp\left(\beta \sum_{\{i,j\}\in E} \mathbf{1}_{(z_i=z_j)}\right), \end{aligned} \tag{2}$$

where $\mathbf{1}_{(z_i=z_j)}$ is the indicator function equal to 1 when $z_i = z_j$ and 0 otherwise, $\{i, j\} \in E$ indicates that $\{i, j\}$ is an edge in G, β is some unknown scalar parameter, and the π_ks are weights defined for every $k \geq 0$ as

$$\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l<k}(1 - \tau_l),$$

where $\boldsymbol{\tau}^\top = (\tau_0, \tau_1, \ldots)$ is a sequence of independent, identically distributed (i.i.d.) random variables with distribution $\mathrm{Beta}(1, \alpha)$. The parameter α is a hyperparameter, which follows a gamma distribution:

$$\alpha | s_1, s_2 \sim \mathcal{G}(s_1, s_2) \ ,$$

while each parameter λ_k in (1) is also distributed according to a gamma distribution:

$$\lambda_k | a_k, b_k \sim \mathcal{G}(a_k, b_k) \ .$$

The characterisation of the π_ks corresponds to a stick-breaking construction (see Lemma 3.4 in Ghosal & Van der Vaart 2017 and Sethuraman 1994, for details) and guarantees that $\sum_{k=0}^{\infty} \pi_k = 1$, which in turn ensures that the distribution in (2) is a valid Markov field (see Lü, Arbel & Forbes 2020, for details). The complete hierarchical model can thus be stated, for $\boldsymbol{z} = \{z_1, \ldots, z_J\}$ and $k = 0, 1, \ldots$, as follows:

$$\begin{aligned}
&\lambda_k | a_k, b_k \sim \mathcal{G}(a_k, b_k), \\
&\alpha | s_1, s_2 \sim \mathcal{G}(s_1, s_2), \\
&\tau_k | \alpha \sim \mathcal{B}(1, \alpha), \\
&\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1}(1 - \tau_l), \\
&p(\boldsymbol{z} | \boldsymbol{\tau}, \beta) \propto \prod_{j \in J} \pi_{z_j}(\boldsymbol{\tau}) \ \exp\left(\beta \sum_{\{i,j\} \in E} \mathbf{1}_{(z_i = z_j)}\right), \\
&y_j | z_j; \boldsymbol{\Lambda}, N_j \sim \mathcal{P}(y_j; \lambda_{z_j} N_j), \quad \text{for each } j \in J.
\end{aligned}$$

3. Inference using variational approximation

In what follows, we propose an adaptation of the VBEM procedure from Lü, Arbel & Forbes (2020) to Poisson emission probabilities.

The observed counts are denoted by $\boldsymbol{y} = \{y_j, j \in J\}$ and the normalising variables are denoted by $\boldsymbol{N}_J = \{N_j, j \in J\}$. The set of parameters to estimate divides into

two subsets, $\boldsymbol{\Phi}^\top = (\beta, s_1, s_2, \boldsymbol{a}^\top)$ with $\boldsymbol{a} = \{a_k, b_k, k = 1, \ldots\}$, which are unknown but fixed parameters; and $\boldsymbol{\Theta}^\top = (\boldsymbol{\alpha}^\top, \boldsymbol{\Lambda}^\top)$, which are random parameters. The hierarchical representation of the model above induces additional latent variables $(\boldsymbol{z}^\top, \boldsymbol{\tau}^\top)$. There is no formal difference in the treatment of random parameters and latent variables, but it is standard to distinguish between them. The term latent variables usually refers to variables whose number increases with the number of observations, while parameters are usually of fixed dimension. Parameters $\boldsymbol{\Phi}$ are estimated by an empirical Bayes principle:

$$\hat{\boldsymbol{\Phi}} = \arg\max_{\boldsymbol{\Phi}} p(\boldsymbol{y}|\boldsymbol{\Phi}) = \arg\max_{\boldsymbol{\Phi}} \int p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta}|\boldsymbol{\Phi})\, d\boldsymbol{z}\, d\boldsymbol{\tau}\, d\boldsymbol{\Theta}.$$

The random parameters and latent variables are handled via their posterior density $p(\boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta}|\boldsymbol{y}, \boldsymbol{N}_J, \hat{\boldsymbol{\Phi}})$. This posterior is not available in close-form due to an intractable normalising constant. In this work, we use a variational approximation principle to provide an approximation of the true posterior. More specifically, the posterior is approximated using a member of a family of distributions q that factorise and that are truncated appropriately to exhibit only a finite number of terms. The infinite state space for each z_i is dealt with by choosing a truncation of the state space to a maximum label K (Blei & Jordan 2006).

In practice, this consists of assuming that the variational distribution $q(\boldsymbol{z}) = \prod_{j \in J} q(z_j)$ and that the q_{z_j}s satisfy $q_{z_j}(k) = 0$, for $k \geq K$, and that the variational distribution on $\boldsymbol{\tau}$ also factorises as $q_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \prod_{k=0}^{K-2} q_{\tau_k}(\tau_k)$, with the additional condition that $\tau_{K-1} = 1$. We thus have the variation approximation, below:

$$q(\boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta}) = q_\alpha(\alpha) \prod_{j=1}^{J} q_{z_j}(z_j) \prod_{k=0}^{K-2} q_{\tau_k}(\tau_k) \prod_{k=0}^{K-1} q_{\lambda_k}(\lambda_k)\ . \tag{3}$$

In (3), K does not represent the number of classes that is actually assumed to exist in the data but an upper bound on the number. In practice, the exact value of K is not critical; K has only to be fixed to a value large enough so as to be higher than the maximum expected number of classes.

The variational approximation procedure is justified by the now standard statement (see e.g., Lü, Arbel & Forbes 2020) that for every function $q(\boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta})$, the marginal likelihood is lower-bounded, where the lower bound is the Kullback-Leibler divergence between $p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta}|\boldsymbol{N}_J, \boldsymbol{\Phi})$ and $q(\boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta})$,

$$\ln\, p(\boldsymbol{y}|\boldsymbol{\Phi}) \geq \mathrm{E}_{q(\boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta})} \left[\ln \frac{p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta}|\boldsymbol{N}_J, \boldsymbol{\Phi})}{q(\boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta})} \right]. \tag{4}$$

The derived variational algorithm is then an alternate maximisation of this bound with respect to each factor in $q(\boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta})$ and $\boldsymbol{\Phi}$. Each maximisation step has some explicit

functional expression although from the computational point of view, some steps may require further approximations. Their descriptions at a coarser level in terms of blocks of parameters is as follows. The iteration index is denoted by (r) in the successive update formulas:

- VE-$\boldsymbol{z}$:

$$q_z^{(r)}(\boldsymbol{z}) \quad \propto \quad \exp\left(\mathrm{E}_{q_{\theta,\tau}^{(r-1)}}[\ln\, p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta} | \boldsymbol{N}_J, \boldsymbol{\Phi}^{(r-1)})]\right); \tag{5}$$

- VE-$\boldsymbol{\Theta}, \boldsymbol{\tau}$:

$$q_{\theta,\tau}^{(r)}(\boldsymbol{\Theta}, \boldsymbol{\tau}) \quad \propto \quad \exp\left(\mathrm{E}_{q_z^{(r)}}[\ln\, p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta} | \boldsymbol{N}_J, \boldsymbol{\Phi}^{(r-1)})]\right); \tag{6}$$

- VM-$\boldsymbol{\phi}$:

$$\boldsymbol{\Phi}^{(r)} \quad = \quad \arg\max_{\boldsymbol{\phi}}\, \mathrm{E}_{q_z^{(r)} q_{\theta,\tau}^{(r)}}[\ln\, p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta} | \boldsymbol{N}_J, \boldsymbol{\Phi})]. \tag{7}$$

Some of these steps are not specific to the Poisson emissions model and have similar forms as the ones derived by Lü, Arbel & Forbes (2020), for Gaussian emissions. These steps are briefly recalled below, omitting the iteration index for simplicity.

VE-α step

We have

$$q_\alpha(\alpha) \propto p(\alpha|s_1, s_2)\ \exp\left(\sum_{k=0}^{K-2} \mathrm{E}_{q_{\tau_k}}\Big[\ln p(\tau_k|\alpha)\Big]\right) = \mathcal{G}(\alpha; \hat{s}_1, \hat{s}_2),$$

with

$$\begin{aligned}\hat{s}_1 &= s_1 + K - 1,\\ \hat{s}_2 &= s_2 - \sum_{k=0}^{K-2} \mathrm{E}_{q_{\tau_k}}\Big[\ln(1-\tau_k)\Big],\end{aligned}$$

and

$$\mathrm{E}_{q_{\tau_k}}\Big[\ln(1-\tau_k)\Big] = \psi(\gamma_{k,2}) - \psi(\gamma_{k,1} + \gamma_{k,2}), \tag{8}$$

where $\psi(\cdot)$ is the digamma function and $(\gamma_{k,1}, \gamma_{k,2})$ are the parameters defining q_{τ_k} (see below).

VE-τ_k step

For every $0 \leq k < K$,

$$q_{\tau_k}(\tau_k) = \mathcal{B}(\tau_k; \hat{\gamma}_{k,1}, \hat{\gamma}_{k,2})$$

with

$$\hat{\gamma}_{k,1} = 1 + \sum_{j=1}^{J} q_{z_j}(k) = 1 + n_k,$$

and

$$\hat{\gamma}_{k,2} = \frac{\hat{s}_1}{\hat{s}_2} + \sum_{l=k}^{K-1} n_l,$$

where we have used the notation $n_k = \sum_{j=1}^{J} q_{z_j}(k)$. Note also that $\sum_{k=0}^{K-1} n_k = J$.

VM-β step

The handling of parameter β is particularly nuanced. Standard variational approximations focus on solving the posterior intractability resulting from the combination of prior distributions and likelihoods, both usually tractable. Unfortunately, when dealing with Markov random field priors, the issue does not only arise from the prior/likelihood combination but already from the intractable normalising constant of the Markov prior. This constant cannot be discarded as it depends on β. Thus, the VM-β step has no explicit solution and requires further approximation. The additional approximation we propose consists of approximating the gradient, with respect to β, of the lower bound (4):

$$\sum_{k=0}^{K-1} \sum_{i \sim j} q_{z_j}^{(r)}(k) q_{z_i}^{(r)}(k) - \sum_{k=0}^{K-1} \sum_{i \sim j} \tilde{q}_{z_j}^{(r)}(k|\beta)\ \tilde{q}_{z_i}^{(r)}(k|\beta), \tag{9}$$

with $\tilde{q}_{z_j}^{(r)}(z_j|\beta)$, defined by:

$$\tilde{q}_{z_j}^{(r)}(z_j = k|\beta) \quad = \quad \frac{\exp(\ln \pi_k(\tilde{\boldsymbol{\tau}}) + \beta \sum_{i \in N(j)} q_{z_i}^{(r)}(k))}{\sum_{l=0}^{\infty} \exp(\ln \pi_l(\tilde{\boldsymbol{\tau}}) + \beta \sum_{i \in N(j)} q_{z_i}^{(r)}(l))}, \tag{10}$$

for all $k = 0, \ldots, K-1$, where $N(j)$ is the set of neighbours of j in G and $\tilde{\boldsymbol{\tau}} = \mathrm{E}_{q_\tau^{(r)}}[\boldsymbol{\tau}]$. Note that in the BNP-HMRF setting, the use of $\tilde{\boldsymbol{\tau}}$ is also an additional necessary approximation; details can be found in Lü, Arbel & Forbes (2020). Regarding β, the approximation can be interpreted as the transfer to the Potts prior of the variational approximation used for the Markov posterior.

The remaining steps are specific to Poisson emission distributions.

VE-λ_k step

For every $0 \leq k < K$, from the general result in (6) it follows that

$$\begin{aligned} q_{\lambda_k}(\lambda_k) &\propto \exp\left(\mathrm{E}_{\boldsymbol{q_z}}\left[\ln\left(p(\lambda_k|a_k,b_k)\prod_{j=1}^{J}p(y_j|z_j,\lambda_{z_j},N_j)\right)\right]\right) \\ &\propto p(\lambda_k|a_k,b_k)\exp\left(\sum_{j=1}^{J}q_{z_j}(k)\mathrm{E}_{q_{z_j}}\left[\ln p(y_j|\lambda_k,N_j)\right]\right) \\ &\propto p(\lambda_k|a_k,b_k)\exp\left(\sum_{j=1}^{J}q_{z_j}(k)\left(-N_j\lambda_k + y_j\ln(N_j\lambda_k) - \ln(y_j!)\right)\right) \\ &\propto p(\lambda_k|a_k,b_k)\exp\left(\sum_{j=1}^{J}q_{z_j}(k)\left(-N_j\lambda_k + y_j\ln(\lambda_k)\right)\right). \end{aligned}$$

Thus q_{λ_k} is a gamma density $\mathcal{G}(\lambda_k|\hat{a}_k,\hat{b}_k)$, where

$$\hat{a}_k = a_k + \sum_{j=1}^{J} y_j q_{z_j}(k) \quad \text{and} \quad \hat{b}_k = b_k + \sum_{j=1}^{J} N_j q_{z_j}(k).$$

This is a consequence of the conjugacy property of gamma priors for Poisson likelihood functions.

VE-Z_j step

For $j \in J$, the general result in (5) leads to

$$q_{z_j}(z_j) \propto \exp\left(\mathrm{E}_{q_{\lambda_{z_j}}}\left[\ln p(y_j|\lambda_{z_j},N_j)\right] + \mathrm{E}_{q_\tau}\left[\ln \pi_{z_j}(\boldsymbol{\tau})\right] + \beta\sum_{i\in N(j)} q_{z_i}(z_j)\right), \quad (11)$$

where $N(j)$ is the set of neighbours of j in G, and for $z_j = k$,

$$\mathrm{E}_{q_\tau}\left[\ln \pi_k(\boldsymbol{\tau})\right] = \mathrm{E}_{q_{\tau_k}}\left[\ln \tau_k\right] + \sum_{l=0}^{k-1}\mathrm{E}_{q_{\tau_l}}\left[\ln(1-\tau_l)\right],$$

with

$$\mathrm{E}_{q_{\tau_k}}\left[\ln(\tau_k)\right] = \psi(\gamma_{k,1}) - \psi(\gamma_{k,1}+\gamma_{k,2}),$$

where $\mathrm{E}_{q_{\tau_l}}\left[\ln(1-\tau_l)\right]$ is given by (8).

The term $\mathrm{E}_{q_{\lambda_k}}\left[\ln p(y_j|\lambda_k, N_j)\right]$ is obtained as follows:

$$\begin{aligned}\mathrm{E}_{q_{\lambda_k}}\left[\ln p(y_j|\lambda_k, N_j)\right] &= \mathrm{E}_{q_{\lambda_k}}\left[-N_j\lambda_k + y_j \ln(N_j\lambda_k) - \ln(y_j!)\right] \\ &= -N_j\mathrm{E}_{q_{\lambda_k}}\left[\lambda_k\right] + y_j \ln(N_j) + y_j\mathrm{E}_{q_{\lambda_k}}\left[\ln(\lambda_k)\right] - \ln(y_j!),\end{aligned}$$

where due to the fact that q_{λ_k} is a gamma density $\mathcal{G}(\lambda_k|\hat{a}_k, \hat{b}_k)$ (see above), $\mathrm{E}_{q_{\lambda_k}}\left[\lambda_k\right] = \hat{a}_k/\hat{b}_k$ and $\mathrm{E}_{q_{\lambda_k}}\left[\ln(\lambda_k)\right] = \psi(\hat{a}_k) - \ln(\hat{b}_k)$.

VM-(a_k, b_k) step

For every $0 \leq k < K$, the VM-(a_k, b_k) step is conducted by maximising

$$\mathrm{E}_{q_{\lambda_k}}\left[\ln \frac{p(\lambda_k|a_k, b_k)}{q_{\lambda_k}(\lambda_k)}\right],$$

which is equivalent to minimising the Kullback-Leibler divergence between $p(\lambda_k|a_k, b_k)$ and q_{λ_k}. Since both densities are gamma, the minimum is obtained whenever $p(\lambda_k|a_k, b_k) = q_{\lambda_k}$ and thus $(a_k, b_k) = (\hat{a}_k, \hat{b}_k)$.

VM-(s_1, s_2) step

The VM-(s_1, s_2) step is conducted by maximising

$$\mathrm{E}_{q_\alpha}\left[\ln \frac{p(\alpha|s_1, s_2)}{q_\alpha(\alpha)}\right].$$

By a similar argument as for the VM-(a_k, b_k) step, $(s_1, s_2) = (\hat{s}_1, \hat{s}_2)$, since both densities are in the gamma family.

Finally, to start the iterative procedure, initial values of the parameters have to be provided and the choice generally has an effect on the quality of the final estimates. Here we resort to several runs of the k-Means algorithm with random initial labels and using the ratio Y_j/N_j (referred to as the 'case ratio') as input data. For each run, the final clustering yields a value $\hat{z}_j$ for each z_j, and $q_{z_j}(k) = \mathbf{1}_{(k=\hat{z}_j)}$. Parameters $\hat{\mathbf{\Lambda}}_k$ are initialised as the sample means of observations within initial class k. Initial values of parameters a_k and b_k are defined so that $\mathrm{E}_{q_{\lambda_k}}\left[\lambda_k\right] = \hat{\mathbf{\Lambda}}_k$ and $\mathrm{var}_{q_{\lambda_k}}\lambda_k = \min_{\ell, \hat{\mathbf{\Lambda}}_\ell > 0} \hat{\mathbf{\Lambda}}_\ell$. For each of these possible initial variational parameters from the multiple runs, we keep the value that yields the highest free energy (4). The computation of the free energy is detailed in Appendix I. In this step, initial values are also required for (s_1, s_2, β), which are arbitrarily set to (1.4, 1, 0).

4. Application to traffic crash risk mapping

The results presented here comprise of data from Victoria, Australia. Crash data between 2014 and 2018 were obtained from Victoria's open data directory (see, Truong & Currie 2019). The number of traffic crashes were aggregated at statistical areas level 2 (SA2s), which are medium-sized functional areas within the Australian Statistical Geography Standard (ASGS). The total number of SA2s in Victoria is 458, excluding several SA2s that have no population according to the Australian Bureau of Statistics (ABS) 2016 census. The SA2 scale has the advantage of offering a good compromise between the number of regions and the spatial resolution, thus ensuring both tractability of algorithms and the interpretability of results.

Since an absolute validation of the risk mapping is difficult, we resort to covariates to assess whether risk level classes encode relevant and contrasted characteristics between different classes. In practice, the set of covariates is the same as the set of possible normalising variables. The primary set of covariates is given in the upper part of Table 1. Additional covariates are derived from the primary set of covariates, whenever they appear to make sense as normalising variables. These derivatives are listed in the lower part of Table 1. The road density variable (RnDens) roughly corresponds to the squared ratio of the road length on the edge length of a square region. The VtrFAR19 variable represents the traffic density as opposed to the absolute traffic load VktFAR19. The VtpopFAR19 variable represents the traffic load per inhabitant.

4.1. Method

We distinguish between two kinds of analyses: exploratory analyses and full runs of the model. Exploratory analyses aim at quickly assessing the potential of several variables as normalising factors for the risk. In this context we resort to reduced numbers of random initialisations and iterations (1,000 and 300, respectively), while full runs may require more initialisations and iterations for some variables (see hereinafter). The methodology regarding exploratory analyses is the following:

1. A principal component analysis (PCA) is performed on the variables listed in Table 1. If a group of variables appears to be strongly correlated to the same axis, we keep only one variable in the group and discard all others.
2. For each possible normalising variable N_j, the ratio $(y_j/N_j)_{1\leq j\leq J}$ are computed, quantised into seven bins and represented on a map. Boundaries between bins are defined by sample quantiles. These bins are anticipated to be close to the expected initial partition.

Table 1. Available covariates: primary set (upper part) and additional derived ones (lower part) with their description sometimes truncated.

Code	Description
AREASQ16	Square area in km2
pop16	Population in 2016 (resident)
VktFAR19	Vehicle-Kilometres Travelled on Freeways and Arterial roads
RnNoItcNS	Number of non-signalised intersection (excl. roundabout)
RnNoItcS	Number of signalised intersection
RnNoItcR	Number of roundabout intersection
RnNoItc	Total number of intersection (sum of 3 types)
RnLen	Total length of all road network (km)
RnSZ7080	Length of posted speed 70 - 80 km/h road (km)
RnSZ90100	Length of posted speed 90 - 100 km/h road (km)
RnSZo100	Length of posted speed over 100 km/h road (km)
RnDens	road density ($\mathrm{RnLen}^2/\mathrm{AREASQ16}$)
$\mathrm{PopDens}$	population density ($\mathrm{pop16}/\mathrm{AREASQ16}$)
$\mathrm{VtrFAR19}$	$\mathrm{VktFAR19}/\mathrm{RnLen}$, Vehicle-Kilometres Travelled on Freeways etc. per km of road
$\mathrm{VtpopFAR19}$	$\mathrm{VktFAR19}/\mathrm{pop16}$, Vehicle-Kilometres Travelled on Freeways etc. per inhabitant
PropSign	proportion of signalised intersections over total number of intersections

3. The numbers of crashes are plotted against the normalising variable. If the model was adequately characterising the observed data, for each risk level, the average number of crashes should be a linear function of N_j, so that the 2D plot should exhibit sets of points clustered around a number of lines that pass through the origin. In particular, if there was only one risk level, the whole data set would exhibit a linear structure. Therefore, as a first evaluation, a linear regression model is estimated and the regression line is presented. If needed, non-linear transforms of N_j are considered (exponential, logarithm, monomials) until the figure exhibits linear structures. This step is purely exploratory.
4. The BNP-HMRF is estimated with $K = 10$ and provides for each region its probability to be at each risk level. A segmentation or partition of the graph into K segments is then obtained by assigning each region/node to the level with the highest probability. The segmentation is represented as a K-color map.
5. The 2D plot described in Step 3 is drawn again, now using different colours, where one colour is used for each label. To assess label separation, lines passing through the origin are drawn using slopes $\mathrm{E}_{q_{\lambda_k}}\left[\lambda_k\right]$. These are compared with linear regression models estimated separately for each class by least squares. Label separation is also quantified by label marginal entropies represented on a map in grey scale.

6. If the number of estimated risk classes in the data is less than K, the model is considered as potentially relevant.

In the latter case, VBEM is run again with more initial values and iterations if necessary (up to 5,000 and 600, respectively), again with $K = 10$. If the relative growth of free energy falls under 10^{-5} before the specified maximum number of iterations is reached, the algorithm is considered to have converged and is stopped. Then analyses of variance are performed on each covariate (except those discarded in step 1) to interpret the risk classes and to examine how their values segregate within regions. Tests are performed at significance level 0.01 and p-values are provided, as well as per-label boxplots, for variables that yield p-values below 0.01. If the number of risk classes is K, it may indicate that the BNP-HMRF model has trouble finding well-separated classes and that the number of classes could be arbitrarily high when the number of regions increases. In that case, a model with continuous risk levels could be more appropriate.

The data, code, python notebook and system environments (conda, docker) required to reproduce these analyses and results are available at `https://gitlab.inria.fr/statify_public/anzj-crash/-/tree/master/notebooks`.

4.2. Results

4.2.1. Normalising variable selection

The first plane in the variable space of the PCA is depicted in Appendix II as Figure 12. The following variables: RnNoItc, RnLen, RnSZ7080, RnDens, and RnSZ90100 were highly correlated with AREASQ16 and thus were discarded. The variables RnSZo100 and VtpopFAR19 had low variability over the regions and were thus not considered as informative. Among the remaining variables, we illustrate here our procedure by considering the population size (variable pop16) and the traffic density (variable VtrFAR19) as normalising variables in turn. Travelled distance-related variables and population size have been widely used as direct and indirect crash exposure, respectively, in previous crash risk analyses (Lord & Mannering 2010; Elvik 2014; Truong & Currie 2019). Results with other normalising variables are provided in Appendix II.

4.2.2. Risk with respect to the population size

We now consider the population size of a region as N_j, i.e., risks are clustered according to the impact of population size on the number of crashes. For illustrative purposes, the map obtained from Step 2 of the exploratory analysis is included in Appendix II as Figure 11. The model led to seven clusters, among which four have negligible frequencies (see Figures 1

and 3). The frequency corresponds to the size of the cluster divided by the total number of regions. In the right-hand part of Figure 1, the legend indicates the lower bound of each bin on case ratio, except for the last bin where the upper bound is also provided. A close-up view on Melbourne is available in Appendix II as Figure 9. Here we do not provide any detailed interpretation for 8 regions in four clusters (clusters labelled as 5 to 9) that essentially are outliers (the regions correspond to zones with 6 to 184 inhabitants, where 9 to 209 crashes occur); see Figures 3 and 4 a). The estimate $\hat{\beta} \approx 0.34$ indicates rather low spatial aggregation of clusters.

For a given region j, the marginal state entropy is defined as

$$H[q_{z_j}] = -\sum_{k=0}^{K-1} q_{z_j}(k) \ln q_{z_j}(k)$$

and measures the uncertainty regarding the value of Z_j, which is between 0 and $\ln K$: $H[q_{z_j}] = 0$ means that $q_{z_j}(k_0) = 1$ for some value k_0 while $H[q_{z_j}] = \ln K$ means that q_{z_j} is a uniform distribution. The cumulative marginal entropy is defined as

$$\sum_{j=1}^{J} H[q_{z_j}].$$

Here the cumulative marginal entropy is 20.7, indicating moderate uncertainty regarding the labels in some regions (compared to the other models considered, hereafter), although the marginal state entropy is below 0.2 in the majority of the regions (see Figure 2). Note that due to the model eliminating some states k during the estimation process (if and only if $\max_j q_{z_j}(k) < 0.5$), the remaining labels may not have consecutive indices since they are identified with their initial indexing in $\{0, \ldots, K-1\}$.

We ordered the labels by increasing values of risk levels. It can be seen from Figure 3 that the fitted linear regression lines are in accordance with slopes induced by expected risk levels, indicating well-separated classes. There is however some larger discrepancy between the two lines in risk level 4, where the regression line has a larger slope than the expected risk level due to possible confusions between levels 4 and 2 in some regions.

Risk level 0 is related to peripheral regions that are close to the capital of Melbourne, and enclaves, which are often regional towns or rural centres with substantial residential developments (see Figure 1; high-resolution versions of all maps are available online). These are small regions (Figure 4 b) with high population sizes (Figure 4 a), high absolute traffic (Figure 4 c) and high traffic densities (Figure 4 d). Risk level 2 is related to peripheral and central regions. These are medium-sized zones with medium population sizes, absolute traffic and traffic densities. Risk level 4 is related to far peripheral and hypercentral regions (relative

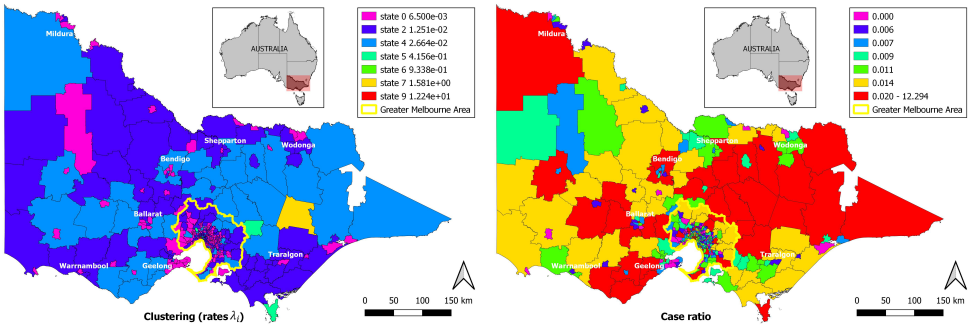

Figure 1. Left-hand part: Risk mapping with respect to population size (variable pop16). Right-hand part: Segmentation using quantiles on ratio. In that part of the figure, the legend indicates the lower bound of each bin on case ratio (sample quantile of order 1/8), except for the last bin where the upper bound is also provided.

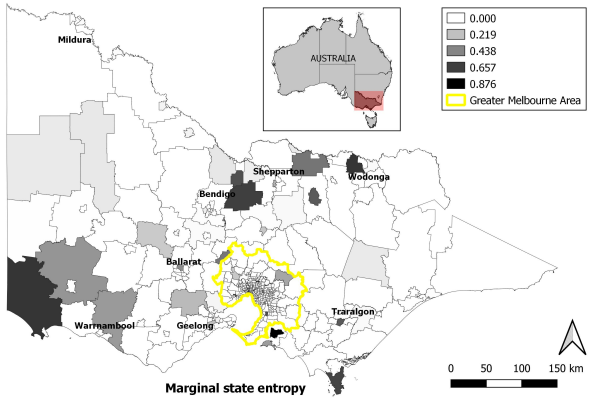

Figure 2. Marginal state entropy for each region regarding risk levels with respect to population size.

to the capital). These are sparsely populated, have varying sizes, with high absolute traffic and traffic densities. The four variables considered in Figure 4 are well discriminated by the risk levels, with ANOVA p-values between 10^{-10} and 10^{-15}, regarding the effects of the classes.

Since regions with higher population sizes have lower risks, it is possible that population has some non-linear effect, which is an avenue for further investigation.

4.2.3. Risk with respect to the traffic density: VtrFAR19 variable

We now set N_j to be the ratio VtrFAR19 of vehicle-kilometers travelled on freeways and arterial roads divided by the total road length in the region, i.e., risks are clustered according to the impact of VtrFAR19 on the number of crashes. N_j can be null in some regions, which leads to a degenerate model, so we set the minimal value to one in VktFAR19

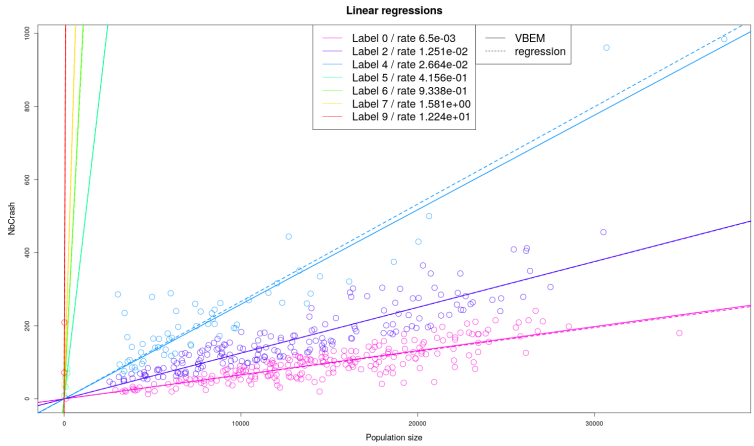

Figure 3. Cluster-wise regressions of crash numbers on population size.

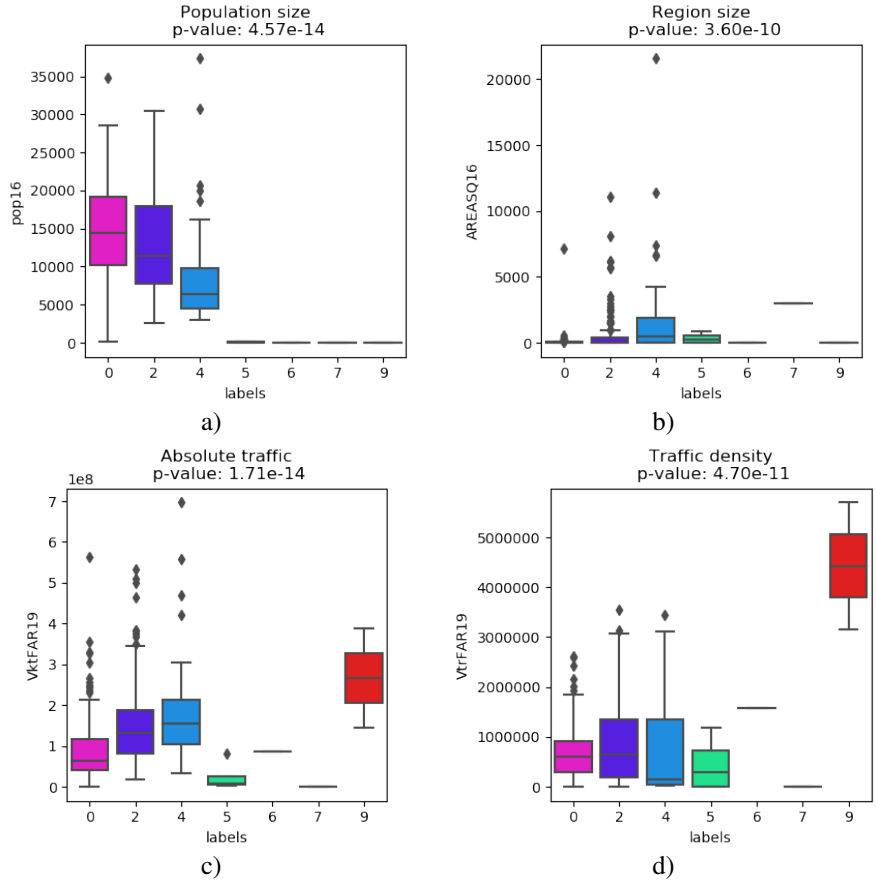

Figure 4. ANOVA of: a) pop16 (population size), b) AREASQ16 (region size), c) VktFAR19 (absolute traffic) and d) VtrFAR19 (traffic density) on risk levels with respect to population size.

when computing the ratio (this happens in the zone of French Island, only). We obtain nine clusters, among which two have negligible frequencies (see Figures 5 and 7). A close-up view on Melbourne is available in Appendix II as Figure 10. The estimate $\hat{\beta} \approx 0.54$ indicates good spatial aggregation of regions. The cumulative marginal entropy is 15.8, indicating mostly low uncertainty regarding risk levels. From Figure 6, the entropy is very close to 0 in a large majority of regions and close to 0.693 (i.e., $\ln 2$) in a small number of regions, which is a value that represents equiprobability between two risk levels. It can be seen from Figure 7 that linear regression lines are in accordance with slopes induced by expected risk levels, indicating well-separated classes. There is however some larger discrepancies between the two values in risk levels 0 and 5, where the regression lines have a larger slopes than the expected risk levels, due to possible confusions between levels 0 and 2, and levels 5 and 6, in some regions.

Risk level 0 is related to regions in the close periphery of the centre (see Figure 5). These are small regions (see Figure 8 b) with high traffic densities, medium population sizes and population densities (see Figures 8 c, a, and d, respectively). Risk level 2 is related to regions in the close periphery of the capital city centre as well as hypercentral regions, and central regions and enclaves. These are small regions with intermediate traffic densities, high population sizes and medium to high population densities. Risk level 3 is related to regions in the close periphery of the centre as well as hypercentral regions, central regions and enclaves. These are small regions with low traffic densities, high population sizes and low population densities. Risk level 5 is related to peripheral regions. These are medium-sized regions with very low traffic densities, population sizes and population densities. Risk level 6 is related to far peripheral regions. These are large regions with very low traffic densities, population sizes and population densities. The four variables considered in Figure 8 are well discriminated by the risk levels, with ANOVA p-values between 10^{-21} and 10^{-69} regarding the effects of classes.

Since regions with higher VtrFAR19 have lower risks, it is possible that VtrFAR19 has some non-linear effect, which we again leave to further investigations.

4.2.4. Combining classes issued from two variables

The analyses performed in Subsections 4.2.2 and 4.2.3 yield different mappings. It would be possible to define new classes as pairs of classes issued from each model. However this would lead to more complex interpretations. Indeed, there are strong dependencies between both classes. A χ^2 independence test rejects the assumption of independence with p-value of 10^{-14}. To visualise associations between classes, a correspondence analysis (CA) was performed. We ignored outlier classes and also class VtrFar19_7 (meaning class 7 in the

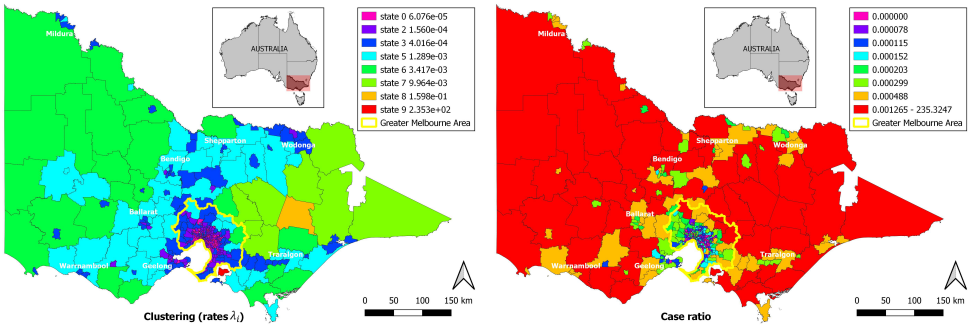

Figure 5. Risk mapping with respect to the traffic density (variable VtrFAR19). Left-hand part: traffic per region. Right-hand part: segmentation using quantiles on ratio.

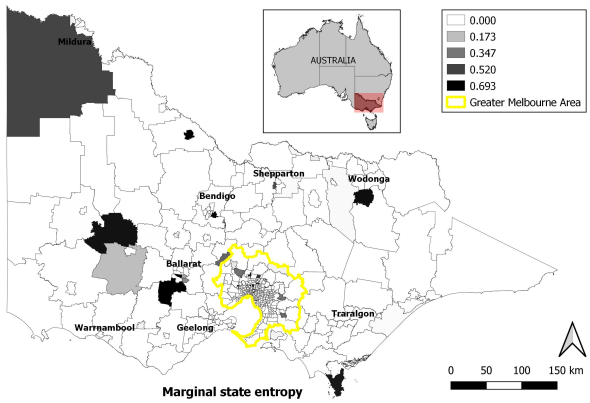

Figure 6. Marginal state entropy for each region regarding risk levels with respect to traffic density VtrFAR19.

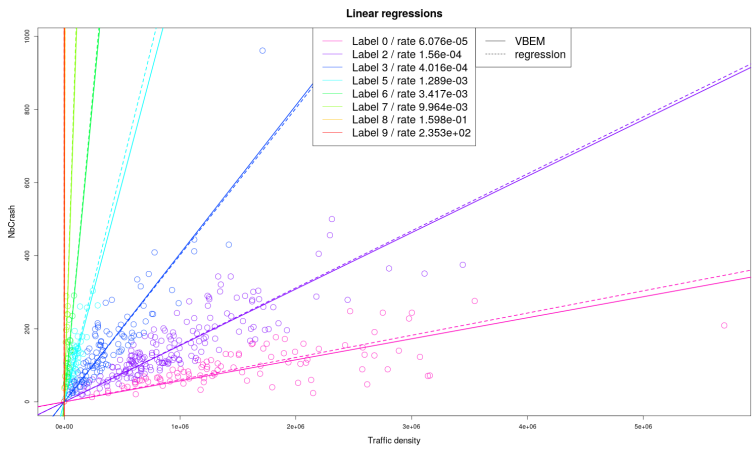

Figure 7. Cluster-wise regressions of crash numbers on traffic density VtrFAR19

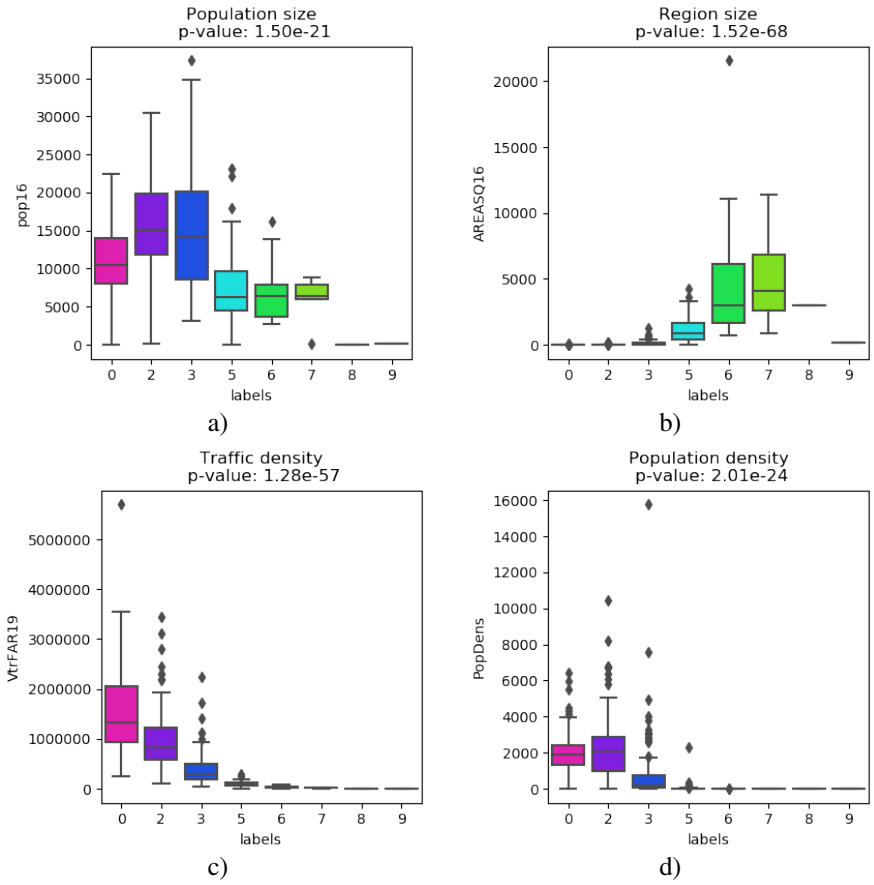

Figure 8. ANOVA of: a) pop16 (population size), b) AREASQ16 (region size), c) VtrFAR19 (traffic density), d) population density on risk levels with respect to traffic density VtrFAR19.

model obtained with normalising risk VtrFar19), which perfectly corresponds to Pop16_4 (higher risk levels in both models). The first CA plane is represented in Appendix II as Figure 13. This shows that labels VtrFar19_2 and VtrFar19_3 are mainly associated into Pop16_0, while Pop16_4 and VtrFar19_5 are strongly associated, VtrFar19_6 is associated with Pop16_2 and VtrFar19_0 is split between Pop16_0 and Pop16_2. As a conclusion, the orders or risk levels in each model are preserved, except for level 6 in VtrFar19.

4.2.5. Risk with respect to other variables

Among the different choices of N_j considered for risk normalisation, the following ones yielded some reduced numbers of classes: VktFAR19, PropSign and PopDens. The associated results are presented in Appendix II. In contrast, RnNoItcNS, RnLen and RnDens

had increasing numbers of classes whenever K increased further beyond 10, indicating some failure in modelling meaningful classes of risks when applying our approach to those variables.

5. Assessment of β estimates yielded by VBEM

The behaviour of VBEM regarding estimation of the β hyperparameter is assessed using simulated data. The data sets were simulated using 3 states and the values of λ_k obtained for the three main clusters in the true data from Subsection 4.2.2: $\lambda_1 = 0.0065, \lambda_2 = 0.013$ and $\lambda_3 = 0.027$. We used the same graph and population values as Subsection 4.2.2. We compare two different settings associated with either $\beta = 0.3$ (estimated value on true data) or $\beta = 0$ (case of spatially independent states). For each set of parameters, 50 simulated data sets are generated, yielding an estimate $\hat{\beta}$ for each of them. The Markov random field was simulated using the SpaceEM^3 software (Vignes et al. 2011). Simulation relies on 100 Gibbs sampling iterations.

5.1. Results when $\beta = 0.3$

The statistics regarding estimation of hyperparameter β and parameters λ_k (estimated as $\mathrm{E}_{q_{\lambda_k}}[\lambda_k]$) are provided in Table 2.

Regarding the number of clusters: the estimated number of clusters was 3 in 31 samples, 4 in 11 samples and between 5 in 8 samples. Whenever the estimated number of clusters was 3, we computed the label discrepancy, defined as the percentage of values in segmentation that match the true labels.

Table 2

parameter	β	λ_1	λ_2	λ_3	label discrepancy
mean	0.55	6.4×10^{-3}	1.2×10^{-2}	2.4×10^{-2}	1.1%
median	0.57	6.4×10^{-3}	1.3×10^{-2}	2.7×10^{-2}	1.0%
standard deviation	0.09	4.2×10^{-3}	2.4×10^{-3}	5.5×10^{-3}	3.7×10^{-3}
minimum	0.31	5.6×10^{-3}	6.9×10^{-3}	1.2×10^{-2}	0.2%
maximum	0.74	7.1×10^{-3}	1.4×10^{-2}	2.8×10^{-2}	1.7%

Although beta was overestimated in our experiments, we observed that this bias did not affect critically the estimation of the others variables and parameters, which remained well estimated.

5.2. Results when $\beta = 0$

For one sample, estimation of β failed since the gradient function defined by (9) was very flat and from a numerical point of view, there was a large set of values of β cancelling the approximated gradient. On the 49 remaining samples, the mean estimated β and the median were 0.03, the standard deviation was 0.04, the minimum was -0.05 and the maximum was 0.14. The test of the null hypothesis $E[\hat{\beta}] = 0$ against the alternative $E[\hat{\beta}] \neq 0$ suggested that β is overestimated (p-value: 10^{-5}).

Regarding estimation of the number of clusters: overestimation in Dirichlet Process independent mixture models is well documented and a post-processing treatment is required to recover consistency properties (Guha, Ho & Nguyen 2021). This behaviour is illustrated in our results: the estimated number of clusters was 3 in 4 samples, between 4 and 5 in 35 samples and between 6 and 9 in 11 samples.

5.3. Conclusion on estimation

The results regarding β estimates on simulated data suggest that this parameter is overestimated, with larger bias when β increases. The number of states is also overestimated, possibly with larger bias when β tends to zero.

6. Conclusion and perspectives

The BNP-HMRF model presented here for risk mapping offers a convenient and fast approach for conducting segmentation of count data regressions indexed by graphs. For example running the 300 iterations required to obtain the results from Subsection 4.2.2 takes between 35 seconds and 40 seconds on a Laptop with an Intel Core 8th i7 8665U CPU with 4 cores, HT, 1.9Ghz, 4.8Ghz Turbo, 8Mo/UHD 620, using hyperthreading. This running time includes selection of the number of classes, which is performed implicitly with state probabilities vanishing when states are not relevant.

The workload required to add new models to our software is rather low, since the specific modifications due to the Poisson assumption can all be embedded into a new class, which is added to the package and by using class inheritance principles. This holds if the added model remains in the exponential family, using conjugate priors.

This models would also offer the possibility to handle missing y_js by introducing new variational factors q_{y_j} in our VBEM approximation. Besides handling incomplete data sets, this would also allow modellers to perform model selection by cross-validation for example.

The proposed model was effective in identifying clusters with distinct risk levels, without the requirement of preselecting the number of clusters K. Furthermore, this is

achieved via a computationally efficient VBEM estimation framework. The proposed BNP model is a significant advancement compared to existing traffic crash mapping approaches, such as the European Road Assessment Program, the BYM model (Besag, York & Mollié 1991), or other spatial CAR models that appear in the road safety literature (Aguero-Valverde & Jovanis 2008; Wang & Kockelman 2013; Truong & Currie 2019), which require arbitrary selection of thresholds for the determination of risk level classes. Like the aforementioned methods, our approach can also be applied via an MCMC implementation, which may permit the use of more structured priors. However, such an implementation incurs an expense of a more computationally intensive estimation procedure, which may not be feasible for large data sets, where a variational approach may be more appropriate. MCMC implementations are available for related models (Orbanz & Buhmann 2008) but none is currently available for our model; thus, we leave the exploration of such procedures to future work.

Regarding the estimation of β using simulated data, we note that the proposed approach is very closed to that in (Forbes & Peyrard 2003). This later mean field approximation provides good results as reported in (Forbes & Peyrard 2003) (and reproduced using the SpacEM3 software). Our BNP setting differs in that it involves a problematic expectation over τ, which can be handled in various ways. In the paper, we propose one such solution but other variants should be investigated. However, note that as in traditional image segmentation, the Potts model is used as a spatial regularization term and is not *per se* a prior model for an image or a partition to be recovered. A typical simulation of a Potts model can be indeed quite far from partitions underlying real data. For this reason, estimating the exact value of β is not the primary target. In practice, although beta is overestimated in our experiments, we observed that it does not affect critically the estimation of the other parameters, which remain well estimated.

This relative insensitivity to the exact β value is not surprising and has been observed and discussed in other context involving HMRF modelling (Pereyra et al. 2013; Forbes & Raftery 1999). Nevertheless, we can propose several ways to prevent this overestimation. A straightforward possibility is to add an exponential prior on β (Chaari et al. 2012). More refined attempts include studying in more details variants of the approximation made over τ to recover performance similar as in traditional parametric HMRF models. These investigations are left for future work.

The number of states also tends to be overestimated, possibly with larger bias when β is close to zero. This point could be addressed in the future by running some post-processing algorithm adapted from (Guha, Ho & Nguyen 2021) on the cluster values.

The proposed model will support the Safe System approach, being adopted in many countries, by identifying crash hot zones for prioritised treatments. Detailed analyses of these clusters showed that regions with higher traffic densities tend to have lower traffic

density-based crash risk levels, while regions with higher population sizes tend to have lower population-based crash risk levels. These findings corroborate well with the traffic safety literature. It is well-established that crash risks tend to decrease with increasing exposure, such as population or the number of road users (safety-in-number effect, Elvik 2014).

Regarding further application to traffic crashes, the model could be extended to consider multiple crash exposure variables. However, the possibility to use several risk-normalising variables leads to multiplying classes and makes their interpretation more difficult. To solve this issue, we could build meta-labels by considering for each region a risk signature. This signature would be a vector of risk levels, each component corresponding to a specific variable N_j. However, this would greatly increase the dimension of the state space to be considered. We could in addition consider either spatial clustering on those signatures (possibly equipped with some metrics) or continuous latent variables obtained with multiple correspondence analysis, introducing continuous latent variables in our model. Moreover, our approach could be compared with clustering results obtained from CAR models, which would require some ranges for grouping CAR random effects as discussed in Section 1.

Ideally, all possibly relevant variables $N_{i,j}$ should be included into a unique multiple regression model such as

$$p(y_j|\lambda_k, N_{1,j}, \ldots, N_{I,j}) = \mathcal{P}(y_j|\lambda_0 + \sum_{i=1}^{I} \lambda_{k,i} N_{i,j})$$

or

$$p(y_j|\lambda_k, N_{1,j}, \ldots, N_{I,j}) = \mathcal{P}\left(y_j|\exp\left[\lambda_0 + \sum_{i=1}^{I} \lambda_{k,i} N_{i,j}\right]\right), \tag{12}$$

and

$$p(y_j|\lambda_k, N_{1,j}, \ldots, N_{I,j}) = \mathcal{P}\left(y_j|\exp\left[u_j + \lambda_0 + \sum_{i=1}^{I} \lambda_{k,i} N_{i,j}\right]\right), \quad u_j \sim \mathrm{N}(0, \tau^2), \tag{13}$$

where (12) and (13) have the respective advantages of enabling non-positive linear predictors and modelling over-dispersion (see also Waller & Carlin 2010). However, such models do not yield explicit VE-steps in the VBEM algorithm. Thus, further approximations would be required, MCMC estimation still being an alternative to overcome this additional difficulty. Although zero-inflated count data did not seem to invalidate our analyses and results, they may have to be considered in other applications, either at the level of Dirichlet Processes (as discussed in Canale et al. 2017) or via alternative emission densities.

To model centrifugal/centripetal effects through a radius r or more generally, the effect of continuous latent variables z_j, we could define the risk $\mathbf{\Lambda}$ as a stochastic, monotonic

function of z_j. Another possibility would be the extension of a scalar β by a function parameterized by the difference in z_j values between two regions.

BNP-HMRF models could also be extended to handle multivariate count data, particularly to address modelling problems in ecology where counts correspond to the number of observed species. The emission densities could thus be replaced by Joint Species Distribution Models (JSDMs), in which spatial dependencies and heterogeneity sources are usually modelled with univariate CARs (see, e.g., Saas & Gosselin 2014) but are ignored in the multivariate count data.

Acknowledgment

We thank the anonymous reviewers for providing us with insightful comments and suggestions.

Appendix I

Free energy computation

Here, we present how to compute the free energy (4). This is mainly used to define a stopping criterion in VBEM, which is why we ignore terms that do not depend on hyper-parameters or variational parameters. The free energy can be decomposed into two terms:

$$\mathrm{E}_{q_{(\boldsymbol{z},\boldsymbol{\tau},\boldsymbol{\Theta})}}\left[\ln \frac{p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\tau},\boldsymbol{\Theta}|\boldsymbol{N}_J,\boldsymbol{\Phi})}{q_{(\boldsymbol{z},\boldsymbol{\tau},\boldsymbol{\Theta})}}\right] = \mathrm{E}_{q_{(\boldsymbol{z},\boldsymbol{\tau},\boldsymbol{\Theta})}}\left[\ln\, p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\tau},\boldsymbol{\Theta}|\boldsymbol{N}_J,\boldsymbol{\Phi})\right] - \mathrm{E}_{q_{\boldsymbol{z}}}\left[\ln\, q_{\boldsymbol{z}}\right] - \mathrm{E}_{q_{\boldsymbol{\tau}}}\left[\ln\, q_{\boldsymbol{\tau}}\right] - \mathrm{E}_{q_{\boldsymbol{\Theta}}}\left[\ln\, q_{\boldsymbol{\Theta}}\right],$$

where the last three terms are entropies and the first term has already been already calculated in the E-step.

Entropy terms

In this section, we provide the expressions for $H[q_z] = -\mathrm{E}_{q_z}[\ln q_z(\boldsymbol{z})]$, $H[q_{\boldsymbol{\tau}}] = -\mathrm{E}_{q_{\boldsymbol{\tau}}}[\ln\, q_{\boldsymbol{\tau}}]$ and $H[q_\theta] = -\mathrm{E}_{q_\theta}[\ln q_\theta(\boldsymbol{\Theta})]$. Firstly,

$$H[q_z] \quad = \quad \sum_{j=1}^{J} H[q_{z_j}] = -\sum_{j=1}^{J}\sum_{k=0}^{K-1} q_{z_j}(k) \ln q_{z_j}(k),$$

where $q_{z_j}(k)$ is given in (11). Next,

$$H[q_\theta] \quad = \quad H[q_\alpha] + \sum_{k=0}^{K-1} H[q_{\lambda_k}]. \tag{14}$$

The first term above can be derived directly from the known entropy expression for a gamma density:

$$H[q_\alpha] = \hat{s}_1 - \ln \hat{s}_2 + \ln \Gamma(\hat{s}_1) + (1 - \hat{s}_1)\psi(\hat{s}_1).$$

The other terms are obtained as follows:

$$-H[q_{\lambda_k}] = \mathrm{E}_{q_{\lambda_k}}\Big[\ln(\lambda_k)|\hat{a}_k, \hat{b}_k\Big] = \psi(\hat{a}_k) - \ln(\hat{b}_k). \tag{15}$$

Moreover,

$$H[q_{\boldsymbol{\tau}}] = \sum_{k=0}^{K-1} H[q_{\tau_k}],$$

where for every k,

$$\begin{aligned} H[q_{\tau_k}] \quad = \ln B(\gamma_{k,1}, \gamma_{k,2}) \quad & -(\gamma_{k,1} - 1)[\psi(\gamma_{k,1}) - \psi(\gamma_{k,1} + \gamma_{k,2})] \\ & -(\gamma_{k,2} - 1)[\psi(\gamma_{k,2}) - \psi(\gamma_{k,1} + \gamma_{k,2})] \end{aligned}$$

and $B(\gamma_{k1}, \gamma_{k2}) = \Gamma(\gamma_{k1})\Gamma(\gamma_{k2})/\Gamma(\gamma_{k1} + \gamma_{k2})$.

E-step terms

The $\mathrm{E}_{q_{(\boldsymbol{z},\boldsymbol{\tau},\boldsymbol{\Theta})}}\left[\ln\, p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta}|\boldsymbol{N}_J, \boldsymbol{\Phi})\right]$ term decomposes into 5 terms:

$$\begin{aligned} \mathrm{E}_{q_{(\boldsymbol{z},\boldsymbol{\tau},\boldsymbol{\Theta})}}\left[\ln\, p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\Theta}|\boldsymbol{N}_J, \boldsymbol{\Phi})\right] &= \sum_{j=1}^{J} \mathrm{E}_{q_{z_j} q_{\boldsymbol{\Lambda}}}[\ln p(y_j|z_j, \boldsymbol{\Lambda})] + \mathrm{E}_{q_\alpha}[\ln p(\alpha|s_1, s_2)] \\ &+ \sum_{k=0}^{K-1} \mathrm{E}_{q_{\tau_k} q_\alpha}[\ln p(\tau_k|\alpha)] + \sum_{k=0}^{K-1} \mathrm{E}_{q_{\lambda_k}}[\ln p(\lambda_k|a_k, b_k)] + \mathrm{E}_{q_z q_\tau}[\ln p(\boldsymbol{z}|\boldsymbol{\tau}, \beta)]. \end{aligned}$$

If the free energy is computed at the end of each VBEM iteration, as per Section 3 (VM-(s_1, s_2) and VM-(a_k, b_k) steps), we have $(s_1, s_2) = (\hat{s}_1, \hat{s}_2)$ and $(a_k, b_k) = (\hat{a}_k, \hat{b}_k)$. Thus, in the free energy $\mathrm{E}_{q_\alpha}[\ln p(\alpha|s_1, s_2)]$ cancels out with $H[q_\alpha]$. Similarly, $\mathrm{E}_{q_{\lambda_k}}[\ln p(\lambda_k|a_k, b_k)]$ cancels out with $H[q_{\lambda_k}]$ in (14).

552 $\sum_{j=1}^{J} \mathrm{E}_{q_{z_j} q_{\mathbf{\Lambda}}}[\ln p(y_j|z_j, \mathbf{\Lambda})]$ ***term.***

553 For each $1 \leq j \leq J$,

$$\mathrm{E}_{q_{z_j} q_{\mathbf{\Lambda}}}\left[\ln p(y_j|z_j, \mathbf{\Lambda})\right] = \sum_{k=0}^{K-1} q_{z_j}(k) \mathrm{E}_{q_{\lambda_k^*}}\left[\ln p(y_j|\lambda_k^*)\right].$$

In the case of Poisson emission densties,

$$\begin{aligned}\mathrm{E}_{q_{z_j} q_{\mathbf{\Lambda}}}\left[\ln p(y_j|z_j, \mathbf{\Lambda})\right] &= \sum_{k=0}^{K-1} q_{z_j}(k) \left[y_j \mathrm{E}_{q_{\lambda_k}}\left[\ln(\lambda_k)\right] - N_j \frac{\hat{a}_k}{\hat{b}_k} + y_j \ln(N_j) - \ln(y_j!)\right] \\ &= y_j \sum_{k=0}^{K-1} q_{z_j}(k) \mathrm{E}_{q_{\lambda_k}}\left[\ln(\lambda_k)\right] - N_j \sum_{k=0}^{K-1} q_{z_j}(k) \frac{\hat{a}_k}{\hat{b}_k} - y_j \ln(N_j) - \ln(y_j!), \qquad (16)\end{aligned}$$

554 with $\mathrm{E}_{q_{\lambda_k}}\left[\ln(\lambda_k)|\hat{a}_k, \hat{b}_k\right]$ given in (15). Since the last two terms in (16) depend on the data
555 only, they do not need to be computed to monitor the convergence of the VBEM algorithm.

556 $\mathrm{E}_{q_{\tau_k} q_\alpha}[\ln p(\tau_k|\alpha)]$ ***term.***

557 Using the expression of a Beta distribution cross-entropy, it follows that

$$\mathrm{E}_{q_{\tau_k} q_\alpha}[\ln p(\tau_k|\alpha)] = \psi(\hat{s}_1) - \ln \hat{s}_2 + \left(\frac{\hat{s}_1}{\hat{s}_2} - 1\right)[\psi(\gamma_{k_2}) - \psi(\gamma_{k1} + \gamma_{k2})].$$

558 $\mathrm{E}_{q_z q_\tau}[\ln p(\boldsymbol{z}|\boldsymbol{\tau}, \beta)]$ ***term.***

559 This term cannot be computed exactly due to the intractable normalising constant $\mathcal{K}$ in
560 the Potts model expression:

$$\ln p(\boldsymbol{z}|\boldsymbol{\tau}, \beta) \quad = \quad \sum_{j=1}^{J} \ln \pi_{z_j}(\boldsymbol{\tau}) + \beta \sum_{i \sim j} \mathbf{1}_{(z_i = z_j)} - \ln \mathcal{K}(\beta, \boldsymbol{\tau}).$$

561 The first two terms can be computed easily but the last one requires approximation. As for the
562 estimation of β, we can use a mean-field like approximation and approximate at each iteration
563 the true value of $\mathcal{K}(\beta, \boldsymbol{\tau})$ by $\tilde{\mathcal{K}}(\beta, \boldsymbol{\tau})$, the normalising constant of $\tilde{q}_z$, defined in (10):

$$\tilde{\mathcal{K}}(\beta, \boldsymbol{\tau}) = \prod_{j=1}^{J} \left(\sum_{l=1}^{\infty} \exp(\ln \pi_l(\boldsymbol{\tau}) + \beta \sum_{i \in N(j)} q_{z_i}^{(r)}(l)) \right),$$

564 where as in Section 3, $\boldsymbol{\tau}$ in the above expression could be replaced by $\mathrm{E}_{q_\tau}[\boldsymbol{\tau}]$ to avoid the
565 dependence in a random $\boldsymbol{\tau}$. However, this would correspond to a zeroth order approximation

and we can do better with a first order approximation as follows:

$$\mathcal{K}(\beta, \boldsymbol{\tau}) \approx \tilde{\mathcal{K}}(\beta, \boldsymbol{\tau})\ \exp(\mathrm{E}_{\tilde{q}_z}[V(\boldsymbol{z}; \boldsymbol{\tau}, \beta) - \tilde{V}(\boldsymbol{z}; \boldsymbol{\tau}, \beta)]), \tag{17}$$

where $V(\boldsymbol{z}; \boldsymbol{\pi}, \beta)$ is defined as

$$V(\boldsymbol{z}; \boldsymbol{\pi}, \beta) = \sum_{j=1}^{N} \log \pi_{z_j} + \beta \sum_{i \sim j} \mathbf{1}_{(z_i = z_j)}$$

and similarly,

$$\tilde{V}(\boldsymbol{z}; \boldsymbol{\tau}, \beta) = \sum_{j=1}^{N} \left(\ln \pi_{z_j}(\boldsymbol{\tau}) + \beta \sum_{i \in N(j)} q_{z_i}^{(r)}(z_j) \right).$$

Note that all terms are tractable:

$$\mathrm{E}_{\tilde{q}_z}[V(\boldsymbol{z}; \boldsymbol{\tau}, \beta) - \tilde{V}(\boldsymbol{z}; \boldsymbol{\tau}, \beta)] = \beta \sum_{j=1}^{N} \sum_{k=0}^{K-1} \tilde{q}_{z_j}(k|\boldsymbol{\tau}, \beta) \left(\frac{\sum_{i \in N(j)} \tilde{q}_{z_i}(k|\boldsymbol{\tau}, \beta)}{2} - \sum_{i \in N(j)} q_{z_i}^{(r)}(k)) \right).$$

Then,

$$\begin{aligned}
&\mathrm{E}_{q_z q_\tau}[\ln \mathcal{K}(\beta, \boldsymbol{\tau})] \approx \mathrm{E}_{q_\tau}[\ln \tilde{\mathcal{K}}(\beta, \boldsymbol{\tau})] + \mathrm{E}_{q_\tau}[\mathrm{E}_{\tilde{q}_z}[V(\boldsymbol{z}; \boldsymbol{\tau}, \beta) - \tilde{V}(\boldsymbol{z}; \boldsymbol{\tau}, \beta)]] \\
&\approx \sum_{j=1}^{N} \mathrm{E}_{q_\tau}[\ln(\sum_{l=1}^{\infty} \exp(\ln \pi_l(\boldsymbol{\tau}) + \beta \sum_{i \in N(j)} q_{z_i}^{(r)}(l)))] \\
&+ \beta \sum_{j=1}^{N} \sum_{k=0}^{K-1} \mathrm{E}_{q_\tau} \left[\tilde{q}_{z_j}(k|\boldsymbol{\tau}, \beta) \left(\frac{\sum_{i \in N(j)} \tilde{q}_{z_i}(k|\boldsymbol{\tau}, \beta)}{2} - \sum_{i \in N(j)} q_{z_i}^{(r)}(k)) \right) \right].
\end{aligned}$$

Appendix II

Additional figures and results on crash data analysis

Here, we present further figures and analyses obtained by using different covariates in our model.

Figure 9 is a close-up on the map in Figure 1 highlighting risk mapping in the regions of Melbourne. Similarly, Figure 10 is a close-up on the map in Figure 5

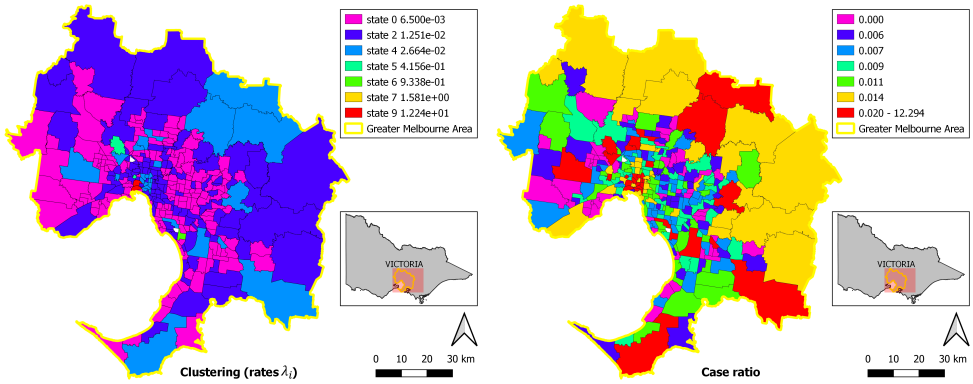

Figure 9. Left-hand part: Risk mapping with respect to population size (variable pop16) in Melbourne. Right-hand part: Segmentation using quantiles on ratio.

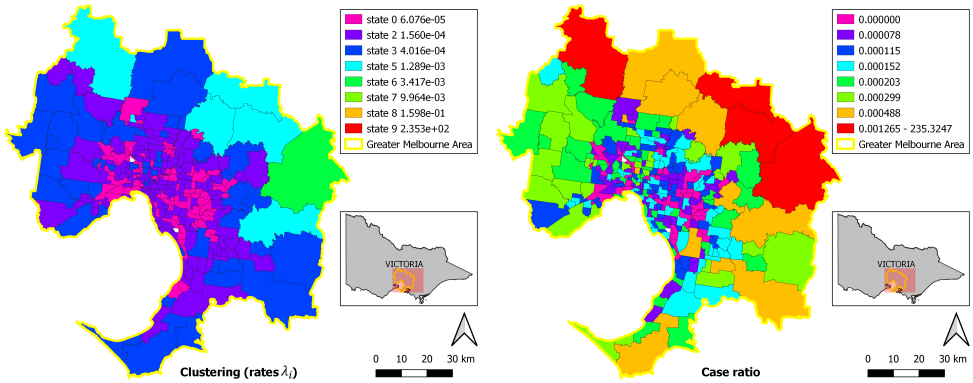

Figure 10. Left-hand part: Risk mapping with respect to the traffic density (variable VtrFAR19) in Melbourne. Right-hand part: Segmentation using quantiles on ratio.

574 Figure 11 depicts the map obtained from Step 2 of the exploratory analysis: population
575 size and number of crashes per region together with the segmentation obtained by using 7
576 quantiles on the ratio of the 'number of crashes on population size'.

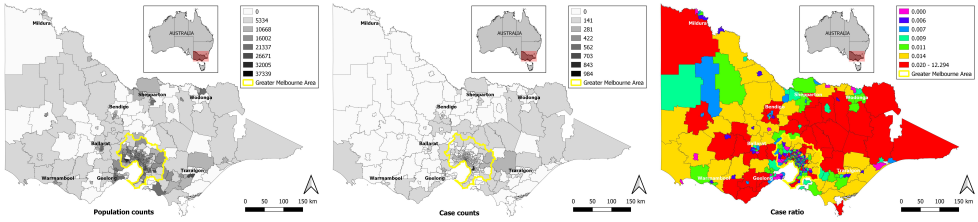

Figure 11. Maps of data and risk ratio used in exploratory using population size (variable pop16). Left: population size per region. Middle: number of crashes per region. Right: segmentation using quantiles on ratio.

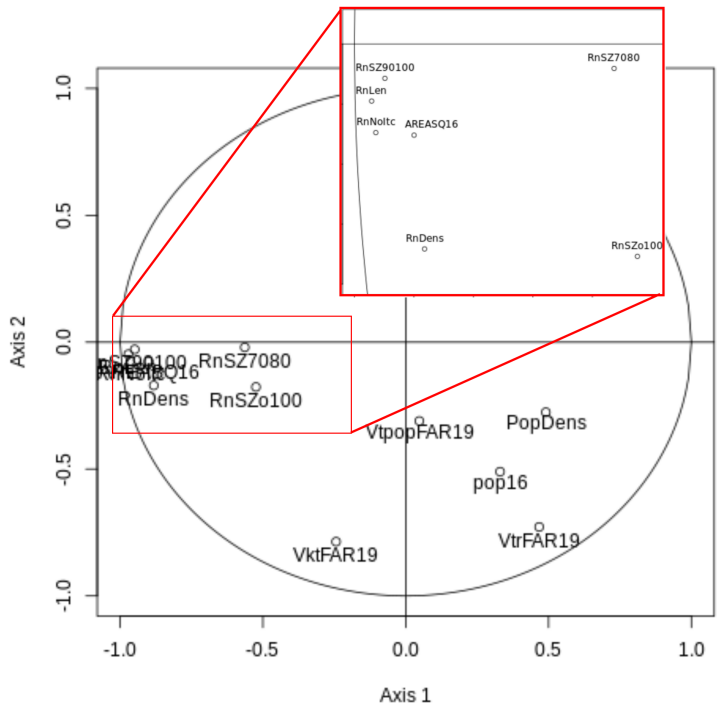

Figure 12. Principal component analysis on covariates: first plane in the variable space. The rectangle framed in red has been zoomed in for the sake of readability.

Figure 12 depicts the first plane in the variable space of the PCA performed on the following variables: RnNoItc, RnLen, RnSZ7080, RnDens, RnSZ90100, AREASQ16, RnSZo100, VtpopFAR19, PopDens, pop16, VktFAR19 and VtrFAR19.

Figure 13 depicts the first plane of the correspondence analysis between the labels obtained using Pop16 (population size) and VtrFar19 (traffic density) as normalising variables.

The following subsections reproduce similar analyses as those that appear in Subsections 4.2.2 and 4.2.3 but using other variables to normalise risks of crashes.

Risk with respect to the traffic: VktFAR19 variable

We now set N_j as the vehicle-kilometres travelled on freeways and arterial roads, i.e., risks are clustered according to the impact of vehicle-kilometres travelled on the number of crashes. The risk appears as somewhat spatially homogeneous and applying the model directly does not yield several well-separated clusters. We consider the transform $p(y_j|z_j = k; \mathbf{\Lambda}, N_j) = \mathcal{P}(y_j; \lambda_k \sqrt{N_j})$. Here, N_j can be null in some regions, which leads

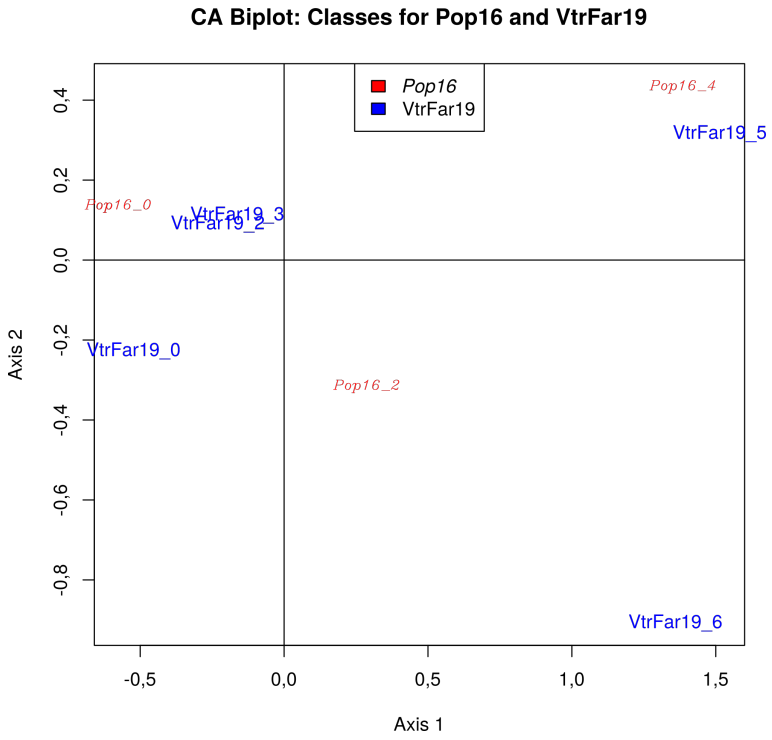

Figure 13. Correspondence analysis between the classes associated with the Pop16 (population size) and VtrFAR19 (traffic density) models: first principal plane.

to a degenerate model, so we set the minimal value to one. We obtain three clusters, all of which are well represented (see Figure 14). The estimate $\hat{\beta} \approx 0.42$ indicates a moderate spatial aggregation of regions. The cumulative marginal entropy is 32.6, indicating some rather high uncertainty regarding risk levels in some regions. It can be seen from Figure 15 that linear regression lines are in accordance with slopes induced by expected risks levels, indicating well-separated classes. There is however some larger discrepancies between the two quantities in risk level 8, where the regression line has a lower slope than the expected risk level, due to possible confusions between levels 8 and 4 in some regions.

Risk level 1 is related to peripheral far west and northern regions (see Figure 14). These are regions with low traffic densities, population sizes and population densities (see Figure 16). Risk level 4 is related to regions in the close periphery of the capital city centre and far east regions. These have intermediate traffic densities, population sizes and population densities. Risk level 8 is related hypercentral regions and those at the close east periphery. These have high traffic densities, population sizes and population densities.

We remark that VktFAR19 has no effect on the labels (p-value 0.69), which seems to confirm the linear relationship between VktFAR19 and crash numbers (cluster-wise).

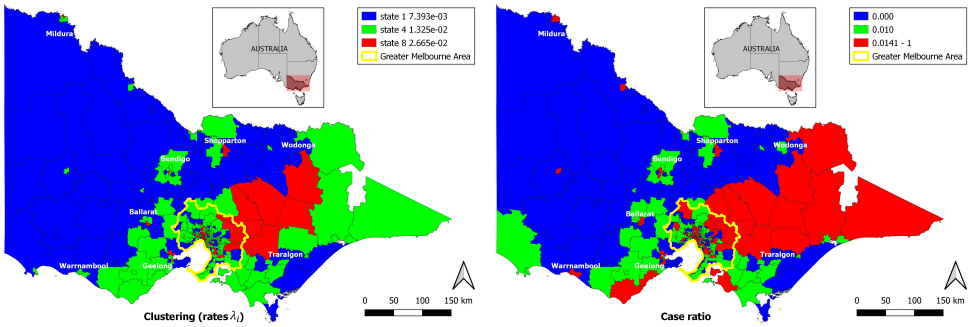

Figure 14. Risk mapping with respect to the traffic (variable VktFAR19). Left-hand part: traffic per region. Right-hand part: segmentation using quantiles on ratio.

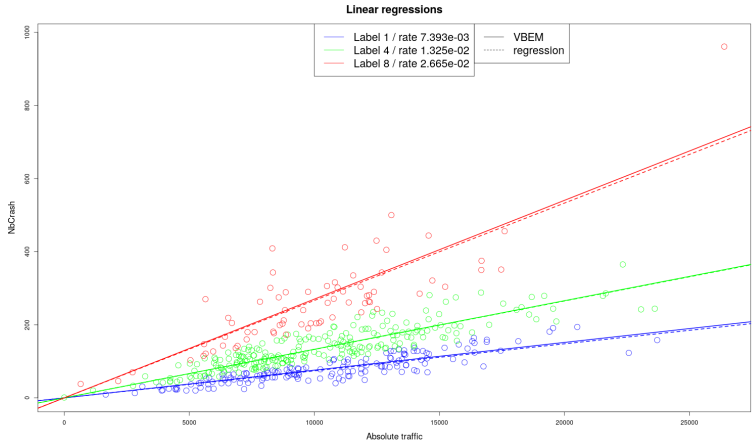

Figure 15. Cluster-wise regressions of crash numbers on VktFAR19.

Risk with respect to the proportion of signalised intersections: PropSign variable

The number of signalised intersections (including roundabouts) was generally quite low and the number of non-signalised intersections was highly correlated with the number of intersections (Pearson correlation: 0.99988). However, the ratio r of the number of signalised intersections on total number of intersections what somehow variable, so we could use it as the normalising variable. We subtracted the minimal value (rounded to the closest multiple of 0.1, here 0.8) and considered $N_j = (r - 0.8)^{1.5}$. In this scenario, we consider that risks are clustered according to the impact of relative increase of the proportion of signalised intersections with respect to the minimal ratio on the number of crashes. We obtained nine clusters, among which four have negligible frequencies (see Figure 7). The estimate $\hat{\beta} \approx 0.27$ indicates low spatial aggregation of regions. The cumulative marginal entropy is 15.8, indicating mostly low uncertainty regarding risk level. Labels are mostly explained

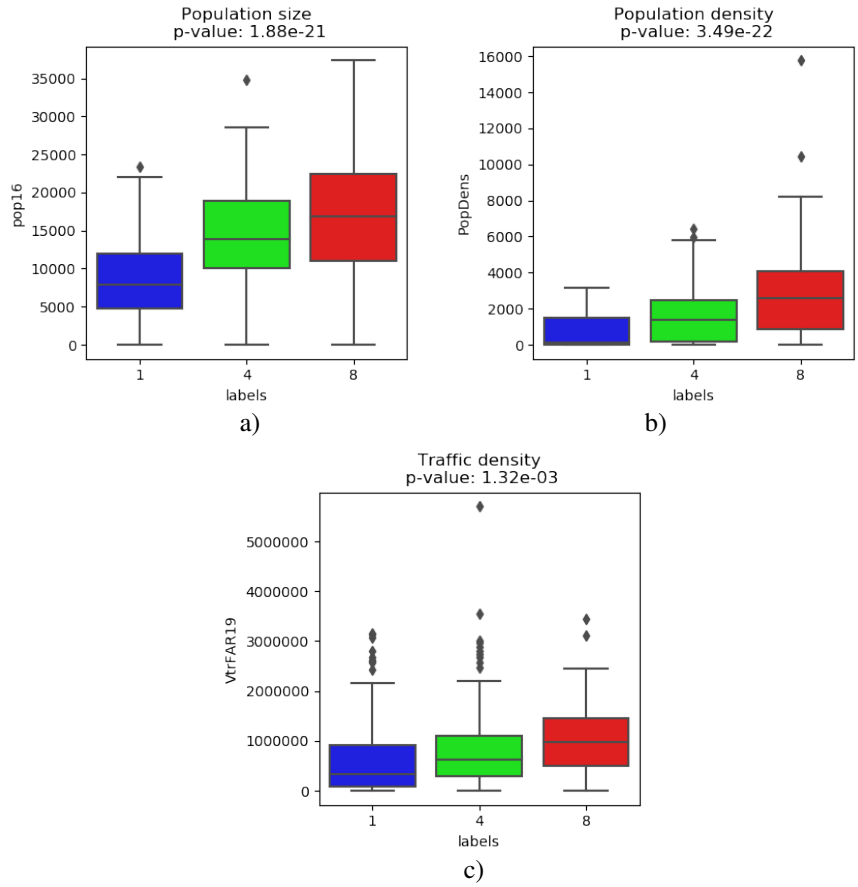

Figure 16. ANOVA of: a) pop16 (population size), b) PopDens (population density) and c) VtrFAR19 (traffic density) on risk levels with respect to absolute traffic intensity VktFAR19.

by centripetal gradients. It can be seen from Figure 18 that the accordance between linear regression lines and slopes induced by expected risks levels is moderate, indicating possible confusions between classes 3 versus 4, 4 versus 5 and 5 versus 6.

Risk levels 5, 6, 7 and 9 apply to a very small numbers of regions (with small values of N_j, see Figure 17) and can be neglected. Risk level 0, 1 and 2 apply to far peripheral regions with low absolute traffics, traffic densities, population sizes and population densities (with an increasing risk regarding those four variables). Risk level 3 is less peripheral than 0, 1, 2 with higher absolute traffics, traffic densities, population sizes and population densities. Risk level 4 applies to more central regions than 3, with slightly higher population densities but somewhat lower population sizes and traffic.

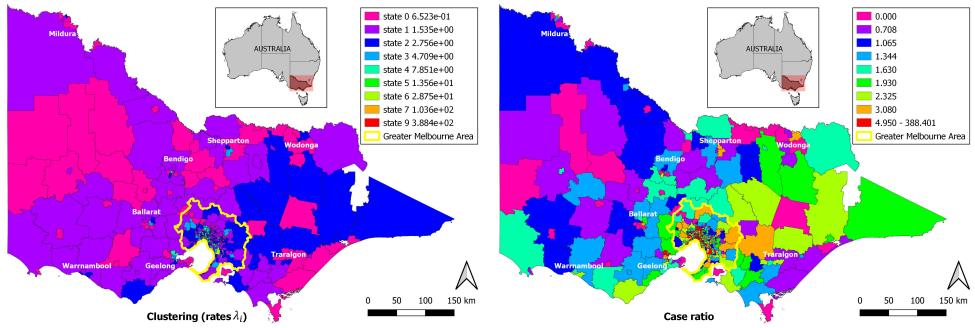

Figure 17. Risk mapping with respect to the ratio of signalised on total number of intersections (variable PropSign). Left-hand part: ratio per region. Right-hand part: segmentation using quantiles on ratio.

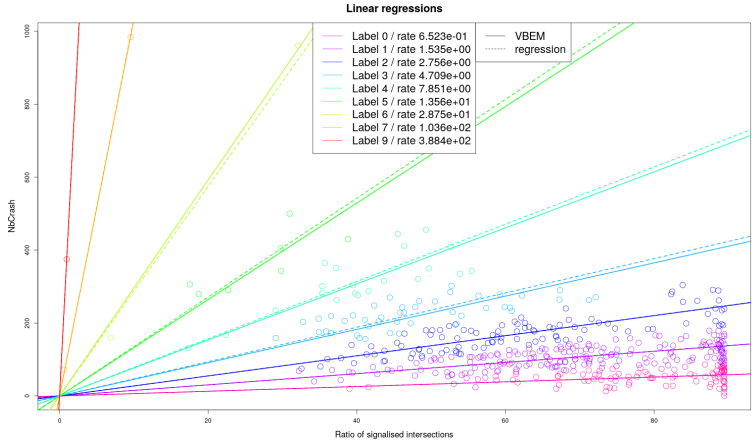

Figure 18. Cluster-wise regressions of crash numbers on the ratio of signalised on total number of intersections.

Risk with respect to the population density: PopDens

We now set N_j as the population density PopDens, i.e., risks are clustered according to the impact of PopDens on the number of crashes. We consider the transform $p(y_j|z_j = k; \mathbf{\Lambda}, N_j) = \mathcal{P}(y_j; \mathbf{\Lambda}, \sqrt{N_j})$. We obtain nine clusters, among which three have negligible frequencies (see Figure 21). The estimate $\hat{\beta} \approx 0.45$ indicates moderate spatial aggregation of regions. The cumulative marginal entropy is 17.7, indicating moderate uncertainty regarding risk level. It can be seen from Figure 21 that the accordance between linear regression lines and slope induced by expected risks levels is moderate, indicating possible confusions between classes 2 versus 3, 3 versus 4 (mainly), 4 versus 5 and 5 versus 6.

Risk levels 7 to 9 apply to a very small numbers of regions (hypercentral or hyperperipheral, with small N_j, see Figure 20) and can be neglected. Risk levels 0 to 6 are

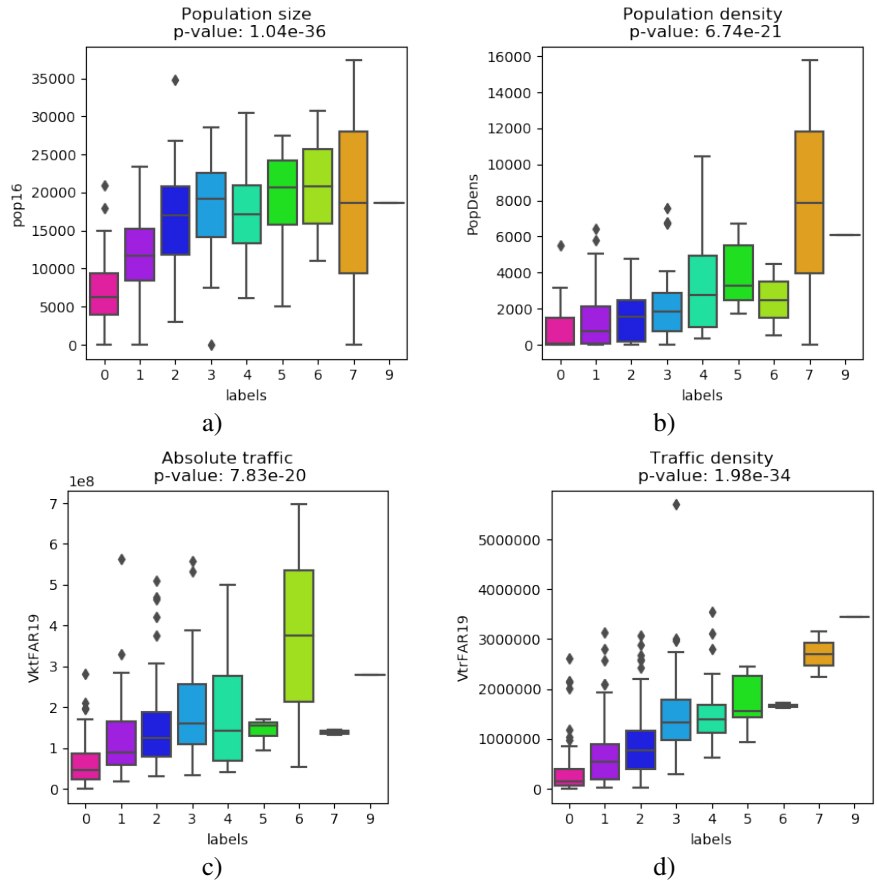

Figure 19. ANOVA of: a) pop16 (population size), b) PopDens (population density), c) VktFAR19 (absolute traffic intensity) and d) VtrFAR19 (traffic density) on risk levels with respect to the proportion of signalised intersections.

mostly explained by centrifugal gradients: from small central regions, with high population densities, low absolute traffics but high traffic densities to large peripheral regions, with low population densities, high absolute traffics but low traffic densities (see Figure 22). Enclaves mostly have risk levels 2 or 0, while the risks of their surrounding regions are higher.

We remark that if a region has low VktFAR19 and high VtrFAR19, it means that it has low absolute traffic but dense traffic, when considering the total road network length.

Since regions with higher PopDens have lower risks, it is possible that PopDens has some non-linear effect, which is a point to be further investigated.

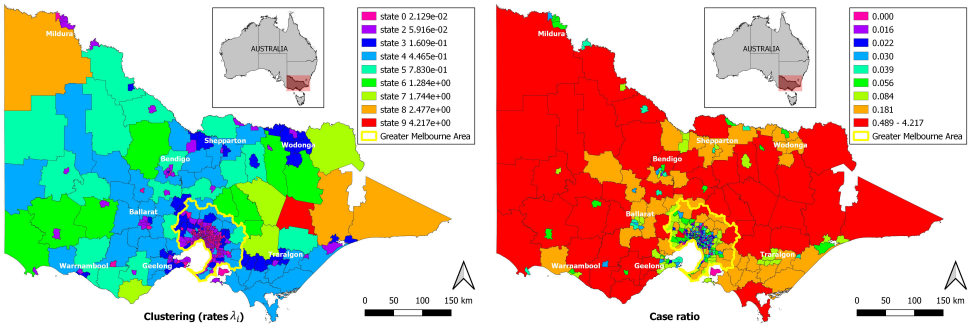

Figure 20. Risk mapping for variable PopDens. Left-hand part: population density per region. Middle part: assigned risk level entropy. Right-hand part: segmentation into a finite number of risk levels.

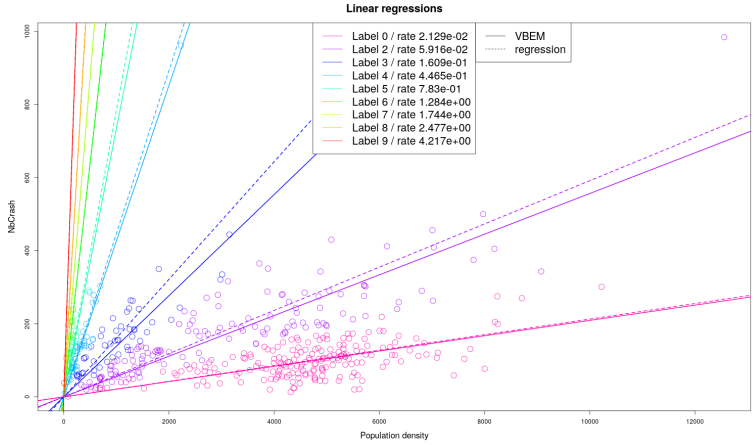

Figure 21. Cluster-wise regressions of crash numbers on PopDens.

References

ABRIAL, D., CALAVAS, D., JARRIGE, N. & DUCROT, C. (2005). Spatial heterogeneity of the risk of BSE in France following the ban of meat and bone meal in cattle feed. *Preventive veterinary medicine* **67**, 69–82. doi:10.1016/j.prevetmed.2004.10.004. URL `http://www.ncbi.nlm.nih.gov/pubmed/15698909`.

AGUERO-VALVERDE, J. & JOVANIS, P.P. (2008). Analysis of road crash frequency with spatial models. *Transportation Research Record* **2061**, 55–63.

ALFO, M., NIEDDU, L. & VICARI, D. (2009). Finite mixture models for mapping spatially dependent disease counts. *Biometrical Journal* **51**, 84–97.

BESAG, J., YORK, J. & MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1–59.

BLEI, D.M. & JORDAN, M.I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1**, 121–143.

BÖHNING, D., DIETZ, E. & SCHLATTMANN, P. (2000). Space-time mixture modelling of public health data. *Statistics in Medicine* **19**, 2333–2344.

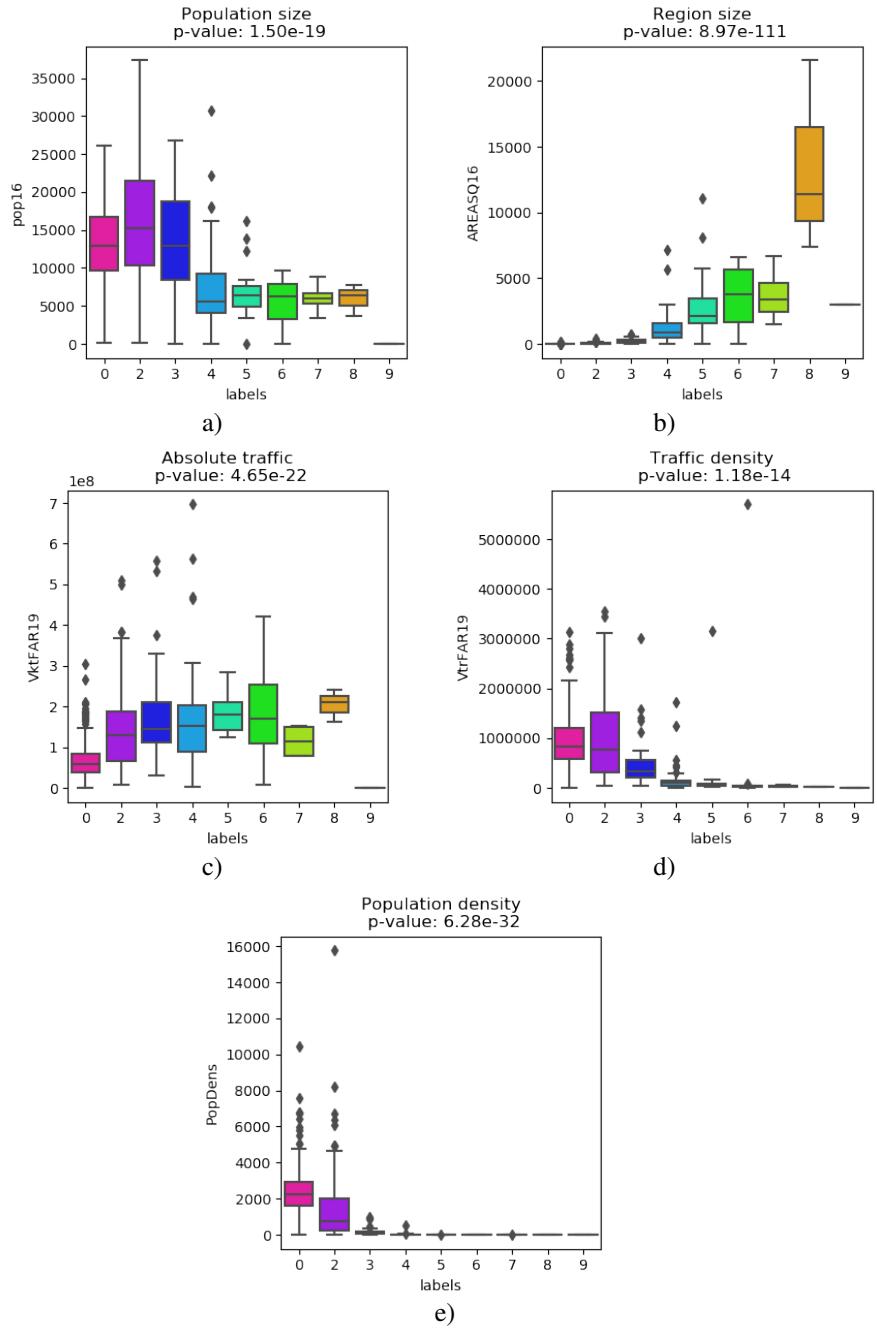

Figure 22. ANOVA of: a) pop16 (population size), b) AREASQ16 (region size), c) VktFAR19 (absolute traffic intensity), d) VtrFAR19 (traffic density) and e) PopDens (population density) on risk levels with respect to population density.

663 CANALE, A., LIJOI, A., NIPOTI, B. & PRÜNSTER, I. (2017). On the pitman-yor process with spike and
664 slab base measure. *Biometrika* **104**, 681–697.
665 CHAARI, L., VINCENT, T., FORBES, F., DOJAT, M. & CIUCIU, P. (2012). Fast joint detection-estimation of
666 evoked brain activity in event-related fMRI using a variational approach. *IEEE transactions on Medical*
667 *Imaging* **32**, 821–837.
668 CHANDLER, D. (1987). *Introduction to modern statistical mechanics*. New York, Oxford: Oxford University
669 Press. URL http://opac.inria.fr/record=b1081336.
670 CLAYTON, D. & BERNADINELLI, L. (1992). Bayesian methods for mapping disease risk. *Geographical*
671 *and Environment Epidemiology: Methods for Small Area Studies, eds. P.Elliot, J.Cuzik, D.English, and*
672 *R.Stern, Oxford, UK:Oxford University Press* , 205–220.
673 ELVIK, R. (2014). *Towards a general theory of the relationship between exposure and risk*. Institute of
674 Transport Economics, Oslo, Norway.
675 ELVIK, R., VAA, T., HOYE, A. & SORENSEN, M. (2009). *The handbook of road safety measures*. Emerald
676 Group Publishing.
677 FERNANDEZ, C. & GREEN, P. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach.
678 *Journal of the Royal Statistical Society: Series B (Methodological)* **64**, 805–826.
679 FORBES, F. & PEYRARD, N. (2003). Hidden Markov random field model selection criteria based on mean
680 field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**, 1089–
681 1101.
682 FORBES, F. & RAFTERY, A.E. (1999). Bayesian morphology: fast unsupervised Bayesian image analysis.
683 *Journal of the American Statistical Association* **94**, 555–568.
684 FRALEY, C. & RAFTERY, A. (2007). Bayesian regularization for normal mixture estimation and model-based
685 clustering. *Journal of Classification* **24**, 155–181.
686 GHOSAL, S. & VAN DER VAART, A. (2017). *Fundamentals of nonparametric Bayesian inference*, vol. 44.
687 Cambridge University Press.
688 GREEN, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model
689 determination. *Biometrika* **82**, 711–732.
690 GREEN, P.J. & RICHARDSON, S. (2002). Hidden Markov models and disease mapping. *Journal of the*
691 *American Statistical Association* **97**, 1–16.
692 GUHA, A., HO, N. & NGUYEN, X. (2021). On posterior contraction of parameters and interpretability in
693 Bayesian mixture modeling. *Bernoulli* **27**, 2159–2188.
694 KNORR-HELD, L. & RASSER, G. (2000). Bayesian detection of clusters and discontinuities in disease maps.
695 *Biometrics* **56**, 13–21.
696 KNORR-HELD, L., RASSER, G. & BECKER, N. (2002). Disease mapping of stage-specific cancer incidence
697 data. *Biometrics* **58**, 492–501.
698 KNORR-HELD, L. & RICHARDSON, S. (2003). A hierarchical model for space-time surveillance data on
699 meningococcal disease incidence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*
700 **52**, 169–183.
701 LAWSON, A., BIGGERI, A., BOEHNING, D., LESAFFRE, E., VIEL, J., CLARK, A., SCHLATTMANN, P.
702 & DIVINO, F. (2000). Disease mapping models: an empirical evaluation. *Statistics in Medicine* **19**,
703 2217–2241.
704 LAWSON, A. & SONG, H. (2010). Bayesian hierarchical modeling of the dynamics of spatio-temporal
705 influenza season outbreaks. *Spatial and Spatio-temporal Epidemiology* **1**, 187–195.
706 LORD, D. & MANNERING, F. (2010). The statistical analysis of crash-frequency data: A review and
707 assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice* **44**,
708 291–305. doi:https://doi.org/10.1016/j.tra.2010.02.001. URL http://www.sciencedirect.
709 com/science/article/pii/S0965856410000376.
710 LÜ, H., ARBEL, J. & FORBES, F. (2020). Bayesian nonparametric priors for hidden Markov
711 random fields. *Statistics and Computing* **30**, 1015–1035. URL https://doi.org/10.1007/

s11222-020-09935-9.

MACNAB, Y. (2010). On Gaussian Markov random fields and Bayesian disease mapping. *Statistical Methods in Medical Research* **20**, 49–68.

MILLER, J.W. & HARRISON, M.T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association* **113**, 340–356.

MOLLIÉ, A. (1996). Bayesian mapping of disease. *Markov Chain Monte Carlo in Practice, eds. W. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall* , 359–379.

MOLLIÉ, A. (1999). Bayesian and empirical Bayes approaches to disease mapping. In *Disease mapping and risk assessment for public health*, eds. A. Lawson, A. Biggeri & D. Bohning. Wiley, pp. 15–29.

ORBANZ, P. & BUHMANN, J. (2008). Nonparametric Bayesian image segmentation. *International Journal of Computer Vision* **77**, 25–45.

PAPADIMITRIOU, E., FILTNESS, A., THEOFILATOS, A., ZIAKOPOULOS, A., QUIGLEY, C. & YANNIS, G. (2019). Review and ranking of crash risk factors related to the road infrastructure. *Accident Analysis & Prevention* **125**, 85 – 97.

PASCUTTO, C., WAKEFIELD, J., BEST, N., RICHARDSON, S., BERNARDINELLI, L., STAINES, A. & ELLIOTT, P. (2000). Statistical issues in the analysis of disease mapping data. *Statistics in Medicine* **19**, 2493–2519.

PEREYRA, M., DOBIGEON, N., BATATIA, H. & TOURNERET, J.Y. (2013). Estimating the granularity coefficient of a Potts-Markov random field within a Markov chain Monte Carlo algorithm. *IEEE Trans. Imag. Process.* **22**, 2385–2397.

RICHARDSON, S., MONFORT, C., GREEN, M., DRAPER, G. & MUIRHEAD, C. (1995). Spatial variation of natural radiation and childhood leukaemia incidence in great britain. *Statistics in Medicine* **14**, 2487–2501.

ROBERTSON, C., NELSON, T., MACNAB, Y. & LAWSON, A. (2010). Review of methods for space-time disease surveillance. *Spatial and Spatio-temporal Epidemiology* **1**, 105–116.

SAAS, Y. & GOSSELIN, F. (2014). Comparison of regression methods for spatially-autocorrelated count data on regularly- and irregularly-spaced locations. *Ecography* **37**, 476–489.

SCHLATTMANN, P. & BÖHNING, D. (1993). Mixture models and disease mapping. *Statistics in Medicine* **12**, 1943–1950.

SETHURAMAN, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica* **4**, 639–650.

STOEHR, J. (2017). A review on statistical inference methods for discrete Markov random fields. *arXiv e-prints* , arXiv:1704.033311704.03331.

THEOFILATOS, A. & YANNIS, G. (2014). A review of the effect of traffic and weather characteristics on road safety. *Accident Analysis & Prevention* **72**, 244 – 256.

TRUONG, L. & CURRIE, G. (2019). Macroscopic road safety impacts of public transport: A case study of Melbourne, Australia. *Accident Analysis & Prevention* **132**, 105270.

VIGNES, M., BLANCHET, J., LEROUX, D. & FORBES, F. (2011). SpaCEM^3: a software for biological module detection when data is incomplete, high dimensional and dependent. *Bioinformatics* **27**, 881–882.

WALLER, L. & CARLIN, B. (2010). Disease mapping. In *Handbook of spatial statistics*, eds. A. Gelfand, P. Diggle, P. Guttorp & M. Fuentes, vol. 2010, chap. 14. Handbook of Modern Statistical Methods. Boca Raton: Chapman & Hall, CRC Press, pp. 217–244.

WANG, Y. & KOCKELMAN, K.M. (2013). A poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. *Accident Analysis & Prevention* **60**, 71–84.